




3 1761 10374543 6



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103745436>

12-001

Government
Publications

Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

June 2012

•

Volume 38

•

Number 1



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access and order this product

This product, Catalogue no. 12-000-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and browse by "Key resource" > "Publications."

This product is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail
Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."



Survey Methodology

A journal
published by
Statistics Canada

June 2012 • Volume 38 • Number 1

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2012

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows:

Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Information Management Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

June 2012

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

MANAGEMENT BOARD

Chairman J. Kovar

Past Chairmen D. Royce (2006-2009)

G.J. Brackstone (1986-2005)

R. Platek (1975-1986)

Members G. Beaudoin

S. Fortier (Production Manager)

J. Gambino

M.A. Hidirolou

H. Mantel

EDITORIAL BOARD

Editor M.A. Hidirolou, *Statistics Canada*

Deputy Editor H. Mantel, *Statistics Canada*

Past Editor J. Kovar (2006-2009)

M.P. Singh (1975-2005)

Associate Editors

J.-F. Beaumont, *Statistics Canada*

J. van den Brakel, *Statistics Netherlands*

J.M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

R. Chambers, *Centre for Statistical and Survey Methodology*

J.L. Eltinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistics Canada*

D. Haziza, *Université de Montréal*

B. Hulliger, *University of Applied Sciences Northwestern Switzerland*

D. Judkins, *Westat Inc*

D. Kasprzyk, *NORC at the University of Chicago*

P. Kott, *National Agricultural Statistics Service*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallée, *Statistics Canada*

P. Lynn, *University of Essex*

D.J. Molec, *National Center for Health Statistics*

G. Nathan, *Hebrew University*

J. Opsomer, *Colorado State University*

D. Pfeffermann, *Hebrew University*

N.G.N. Prasad, *University of Alberta*

J.N.K. Rao, *Carleton University*

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

P. do N. Silva, *Escola Nacional de Ciências Estatísticas*

P. Smith, *Office for National Statistics*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

M. Thompson, *University of Waterloo*

V.J. Verma, *Università degli Studi di Siena*

K.M. Wolter, *Iowa State University*

C. Wu, *University of Waterloo*

W. Yung, *Statistics Canada*

A. Zaslavsky, *Harvard University*

Assistant Editors C. Bocci, K. Bosa, P. Dick, G. Dubreuil, S. Godbout, C. Leon, S. Matthews, Z. Patak, S. Rubin-Bleuer and Y. You, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

Survey Methodology

A Journal Published by Statistics Canada

Volume 38, Number 1, June 2012

Contents

Regular Papers

Sebastian Bredl, Peter Winker and Kerstin Kötschau A statistical approach to detect interviewer falsification of survey data	1
Steven Elliott The application of graph theory to the development and testing of survey instruments	11
G. Hussain Choudhry, J.N.K. Rao and Michael A. Hidirolou On sample allocation for efficient domain estimation	23
Ted Chang Calibration alternatives to poststratification for doubly classified data	31
Paul Knottnerus and Arnout van Delden On variances of changes estimated from rotating panels and dynamic strata	43
Dan Liao and Richard Valliant Variance inflation factors in the analysis of complex survey data	53
Hung-Mo Lin, Hae-Young Kim, John M. Williamson and Virginia M. Lesser Estimating agreement coefficients from sample survey data	63
Jörg Drechsler and Jerome P. Reiter Combining synthetic data with subsampling to create public use microdata files for large scale surveys	73
Balgobin Nandram and Myron Katzoff A hierarchical Bayesian nonresponse model for two-way categorical data from small areas with uncertainty about ignorability	81

Short Notes

Phillip S. Kott Why one should incorporate the design weights when adjusting for unit nonresponse using response homogeneity groups	95
---	----

In Other Journals	101
-------------------------	-----

⑧

A statistical approach to detect interviewer falsification of survey data

Sebastian Bredl, Peter Winker and Kerstin Kötschau¹

Abstract

Survey data are potentially affected by interviewer falsifications with data fabrication being the most blatant form. Even a small number of fabricated interviews might seriously impair the results of further empirical analysis. Besides reinterviews, some statistical approaches have been proposed for identifying this type of fraudulent behaviour. With the help of a small dataset, this paper demonstrates how cluster analysis, which is not commonly employed in this context, might be used to identify interviewers who falsify their work assignments. Several indicators are combined to classify 'at risk' interviewers based solely on the data collected. This multivariate classification seems superior to the application of a single indicator such as Benford's law.

Key Words: Data fabrication; Falsifier; Benford's law; Cluster analysis.

1. Introduction

Whenever data collection is based on interviews, one has to be concerned about data quality. Data quality can be affected by false or imprecise answers of the respondent or by a poorly designed questionnaire, as well as by the interviewer when he or she deviates from the prescribed interviewing procedure. If the interviewer does so consciously, this is referred to as 'interviewer falsification' (Schreiner, Pennie and Newbrough 1988) or 'cheating' (Schräpler and Wagner 2003).

Interviewer falsification can occur in many ways (*cf.* Guterbock 2008). Rather subtle forms consist of surveying a wrong household member or of conducting the survey by telephone when face-to-face interviews are required. The most severe form of falsifying is the fabrication of entire interviews without ever contacting the respective household. In our analysis, we deal with the latter case.

Fabricated interviews can have serious consequences for statistics based on the survey data. Schnell (1991) and Schräpler and Wagner (2003) provide evidence that the effect on univariate statistics might be less severe, provided the share of falsifiers remains sufficiently small and the 'quality' of the fabricated data is high. But even a small proportion of fabricated interviews can be sufficient to cause heavy biases in multivariate statistics. Schräpler and Wagner (2003) find that the inclusion of fabricated data from the German Socio Economic Panel (GSOEP) in a multivariate regression reduces the effect of training on log gross wages by approximately 80 percent, although the share of fabricated interviews was less than 2.5 percent. This indicates the importance of identifying these interviews.

The most common way to identify falsifying interviewers is the reinterview (Biemer and Stokes 1989). In this case, a supervisor contacts some of the households that should have been surveyed to check whether they were actually visited by the interviewer. However, for reasons of expense, it is impossible to reinterview all households participating in a survey (*cf.* Forsman and Schreiner 1991). Therefore, the question arises of how the reinterview sample can be optimized to best detect falsifiers. Generally, it seems useful to select households for reinterview if the interviews were done by an interviewer – identified by characteristics linked to the answers in his interviews – who is more likely than others to be fabricating data. In this context, Hood and Bushery (1997) uses the term 'at risk' interviewer. If reinterview participants are sampled in a two-stage setting, whereby interviewers are selected in the first stage and participants surveyed by those interviewers in the second stage (as recommended by Forsman and Schreiner (1991)) one might oversample the at risk interviewers in the first stage.

In this paper, we demonstrate a purley statistical approach that relies on the data contained in the questionnaires to define a group of at risk interviewers. This is not a new idea; literature provides several examples for this kind of approach (Hood and Bushery 1997; Diekmann 2002; Turner, Gribbe, Al-Tayyip and Chromy 2002; Schräpler and Wagner 2003; Swanson, Cho and Eltinge 2003; Murphy, Baxter, Eyerman, Cunningham and Kennet 2004; Porras and English 2004; Schäfer, Schräpler, Müller and Wagner 2005; Li, Brick, Tran and Singer 2009). However, with the exception of the work of Li *et al.* (2009), the tests conducted in these studies rely on the examination of single indicators derived from the interviewer's data to detect falsifiers. Some studies calculate several indicators but consider them all separately. We combine multiple indicators in cluster

1. Sebastian Bredl, Department of Statistics and Econometrics, Justus-Liebig-University, 35394 Gießen, Licher Straße 64, Germany. E-mail: sebastian.bredl@wirtschaft.uni-giessen.de; Peter Winker, Department of Statistics and Econometrics, Justus-Liebig-University, 35394 Gießen, Licher Straße 64, Germany. E-mail: peter.winker@wirtschaft.uni-giessen.de; Kerstin Kötschau, Hanse Parlament, 22587 Hamburg, Blankeneser Landstrasse 7, Germany. E-mail: kkoetschau@hanse-parlament.eu.

analyses, allowing for a better classification of the potential falsifiers compared to previous approaches. To the best of our knowledge, this procedure is an innovation in the context of identifying interviewers who fabricate data, but has already been employed in other fields in order to detect fraudulent behaviour. The basic idea is that characteristics of fraudulent 'cases' (what a case is depends on the context) feature striking patterns compared to honest cases that can be revealed if those characteristics are jointly considered in a cluster analysis. Murad and Pinkas (1999) try to detect fraud in the telecommunication industry by means of clustering call profiles of clients. A call is characterized by several indicators like calling time or destination of the call. Thiprungsri (2010) clusters group life claims submitted from clients to life insurance companies based on several characteristics of the claims. Claims that form very small clusters are considered to be suspicious. Donoho (2004) uses cluster analysis, among others, to trace patterns in option markets that might indicate insider trading.

We have a small survey dataset available (see subsection 3.1 for a further description of our dataset), which partially consists of falsified data. With a total of 13 interviewers and 250 questionnaires, the size of the dataset is quite limited and it is not clear to what extent our findings can be generalized to larger datasets. However the dataset enables us to demonstrate our approach. The fact that we know which data was collected honestly and which data was fabricated allows for a first evaluation of our approach. It must be stated that this a priori knowledge is no prerequisite to employ the method.

The problem of identifying at risk interviewers was addressed in the 1980s, however, literature on this issue is still scarce. In 1982, the U.S. Census Bureau implemented the Interviewer Falsification Study. Based on the information collected in the context of this study, Schreiner *et al.* (1988) find that interviewers with a shorter length of service are more likely to fabricate data. Hood and Bushery (1997) use several indicators to find at risk interviewers in the National Health Interview Survey (NHIS). For example, they calculate the rate of households that have been labelled ineligible or the rate of households without telephone number per interviewer and compare the rates to census data from the respective area. When large differences occur, the interviewer is flagged and a reinterview is conducted. Detection rates among the flagged interviewers turn out to be higher than those in random reinterview samples. Turner *et al.* (2002) also find interviewers committing data fabrication to indicate telephone numbers less frequently than honest interviewers when examining the Baltimore STD and Behaviour Survey. For the case of computer assisted interviewing, Bushery, Reichert, Albright and Rossiter (1999) and Murphy *et al.* (2004) propose the use of date and

time stamps - the recording of the time and the duration of the interview by the computer - to find suspect interviewers. Those who need a remarkably long or short time to complete the entire questionnaire or certain modules or complete remarkably many questionnaires within a given time period might be flagged as at risk interviewers. Schäfer *et al.* (2005) assume that falsifiers avoid extreme answers when fabricating data. Using data of the GSOEP, the authors calculate the variance of the answers for every question on all questionnaires of an interviewer and sum up all variances. Thanks to other control mechanisms in the GSOEP, falsifiers are known and it turns out that they could be found among the interviewers with the lowest overall variances. Porras and English (2004) use a similar approach and also find falsifiers to produce variances that are smaller to those found in honestly filled questionnaires. Li *et al.* (2009) combine several predictive indicators in a logistic regression model in which the known falsification status of an interview serves as a binary dependent variable. The authors find that reinterview samples that overweight cases with a high probability of being fraudulent according to the logistic regression model identify more cases of actual data fabrication than purely randomly drawn samples. However, it is evident that past reinterview data with known falsification status must be available to conduct the logistic regression.

Further indicators discussed in literature are the number of rare or unlikely response combinations in an interviewer's questionnaires (Murphy *et al.* 2004; Porras and English 2004) and the comparison of household compositions or descriptive statistics in interviewer's questionnaires with the entire sample (Turner *et al.* 2002; Murphy *et al.* 2004).

Another means of detecting fabricated data that has gained a lot of popularity in recent years is Benford's law (Schräpler and Wagner 2003; Swanson *et al.* 2003; Porras and English 2004; Schäfer *et al.* 2005), which will be discussed in section 2, along with its success in detecting fabricated interviews in previous studies. Furthermore, section 2 describes our statistical approach to identify falsifiers. Section 3 presents the data our analysis is based upon as well as our results. The paper concludes with a discussion of our findings.

2. Methods

2.1 Benford's law

When the physicist Frank Benford noticed that the pages in logarithmic tables containing the logarithms of low numbers (1 and 2) were more used than pages containing logarithms of higher numbers (8 and 9), he started to investigate the distribution of leading digits in a wide range

of different types of numbers like numbers on the first page of a newspaper, street addresses or molecular weights (Benford 1938). Benford found that the distribution of the leading non-zero digits could be described by the following formula which has become known as 'Benford's law':

$$\text{Prob}(\text{leading digit} = d) = \log_{10}\left(1 + \frac{1}{d}\right). \quad (1)$$

However, not all series of numbers Benford (1938) investigated seemed to conform to his law. Consequently, the question arose what kind of data can be supposed to produce first digits in line with the law. Discussions of this issue are provided by Hill (1995), Nigrini (1996), Hill (1999) and Scott and Fasli (2001). The detection of financial fraud is a field in which the application of Benford's law has gained much popularity during the recent decade (Nigrini 1996; 1999; Saville 2006). The results of those studies are not relevant in our context. However, it is interesting to note that there seems to be a consensus in literature that monetary values can be supposed to follow Benford's law. Swanson *et al.* (2003) show that the distribution of first digits in the American Consumer Expenditure Survey is close to Benford's distribution.

The basic idea of using Benford's law to detect fabricated data is that falsifiers are unlikely to know the law or to be able to fabricate data in line with it. Therefore a strong deviation of the leading digits from Benford's distribution in a dataset indicates that the data might be faked. Of course, one has to be concerned if the nature of the data is such that it can be supposed to follow Benford's law if it is authentic. Benford's law cannot be applied if the questionnaires do not contain any or contain only very few metric variables.

Schräpler and Wagner (2003) and Schäfer *et al.* (2005) use Benford's law to detect data fabrication in the GSOEP. In both studies, all questionnaires delivered by every single interviewer are combined and checked for whether the distribution of the first digits in the respective questionnaires deviates significantly from Benford's law. This can be done by calculating the χ^2 -statistic:

$$\chi^2_i = n_i \sum_{d=1}^9 \frac{(h_{id} - h_{bd})^2}{h_{bd}} \quad (2)$$

where n_i is the number of leading digits in all questionnaires from interviewer i , h_{id} is the observed proportion of leading digit d in all leading digits in interviewer i 's questionnaires and h_{bd} is the proportion of leading digit d in all leading digits under Benford's distribution. High χ^2 -values indicate a deviation from Benford's distribution and indicate at risk interviewers. Schräpler and Wagner (2003) use different kinds of continuous variables in their analysis, whereas Schäfer *et al.* (2005) restrict theirs to monetary values. In both studies, the critical χ^2 -values are assumed to

be dependent on the sample size n and are consequently adjusted for this parameter. The results obtained look promising. The fit of the leading distribution of first digits to Benford's distribution in the questionnaires of falsifiers (which were already known in advance) is, in general, much worse than for honest interviewers. Thus it seems appropriate to use Benford's law as a means to identify at risk interviewers.

However, when we compared the data of the honest interviewers in our dataset to Benford's distribution, we observed a large deviation for the digit 5. This might be due to rounding of numbers by the respondents. The same problem is mentioned by Swanson *et al.* (2003) and Porras and English (2004) who opt for applying an alternative approach "in the spirit of Benford" (Porras and English 2004, page 4224). We adopt this approach which consists of comparing the distribution of leading digits in the questionnaires of an interviewer to the distribution of first digits in all questionnaires except their own. The χ^2 -value on the interviewer level is calculated as described above but the expected proportion of a digit according to Benford's law h_{bd} is replaced by the proportion of the digit in all other questionnaires. We then use the resulting χ^2 -value as one indicator in the cluster analysis.

With regard to the selection of variables whose first digits are examined, we stick to the approach of Schäfer *et al.* (2005) and include only the first digits of monetary values in the analysis. The survey we are using for demonstration purposes contains monetary values expressed in local currency referring to household expenditures for different items like leasing or buying land, seeds, fertilizer, taxes, and to household income from different sources like agricultural or non agricultural self employment and public or private transfers. Overall we include first digits of 26 different monetary values per interview, ignoring values that were reported to be zero. We then pool first digits of all questionnaires delivered by one interviewer and compare the distribution of first digits to the one for all other interviews according to the method described above. The restriction to monetary values constitutes a clear criterion during the process of selecting data. Furthermore, as mentioned above, financial data is broadly agreed upon to be apt for the analysis with Benford's law. This is important, although we do not ground our analysis on Benford's distribution but on an approach based on it.

2.2 Multivariate analyses

Our idea is to combine several indicators, which we derive directly from the questionnaires of each interviewer and which we suppose to be different for falsifiers and honest interviewers. We do this by means of cluster and discriminant analysis. All indicators are derived on the

interviewer level. This implies that we pool all questionnaires of one interviewer for the analysis, which increases the amount of data on which every single indicator value is based. This should make the indicator values more reliable and less sensitive to outliers. On the other hand, it is obvious that the discriminatory power of interviewer-level indicators decreases as soon as interviewers only fake parts of their assignments. Looking at indicators on the questionnaire level, therefore, seems to be preferable if the amount of data per questionnaire is sufficiently high.

The cluster analysis constitutes the real method of identifying at risk interviewers. The interviewers are clustered in two groups with the intention of obtaining one that contains a high share of falsifiers and another one that contains a high share of honest interviewers. Clustering does not require a priori information on who is fabricating data and who is not. In fact, this is what it is supposed to reveal. Since we know from the outset which interviewer belongs to which group, we can discover whether the cluster analysis identifies the 'true falsifiers' to be at risk. Clearly, the assumption that our approach is able to separate both groups perfectly is not realistic. The idea is rather that we obtain an at risk interviewer cluster exhibiting a higher share of falsifiers compared to the other cluster. If a reinterview is feasible, subsequent reinterview efforts might be focused on interviewers in the at risk cluster.

To judge the performance of the cluster analysis, we consider the number of undetected falsifiers as well as the number of 'false alarms.' Both types of 'errors' incur costs: data of undetected falsifiers is likely to impair the results of further statistical analysis. False alarms incur costs in the sense that an unnecessary effort to reinterview the respective households might be taken or data is unnecessarily removed from the sample. Furthermore, it might be demoralizing for honest interviewers if they see their work being subject to a reinterview, particularly if they are aware of the fact that predominantly the work of at-risk interviewers is picked. How to weight an undetected falsifier compared to a false alarm in a loss function is a highly subjective issue. Generally, it seems reasonable to assign more weight to the former than to the latter.

The discriminant analysis requires knowledge on the falsifiers versus non-falsifiers status of each interviewer before it can be conducted. Therefore, it is not an instrument to detect falsifiers. We use the discriminant analysis to verify our hypotheses on the behaviour of falsifiers, which will be discussed below, and to evaluate how well the employed indicators can separate the two groups.

One of the indicators we use is the χ^2 -value, calculated by comparing the distribution of first digits in the questionnaires of each interviewer with the respective distribution in all other questionnaires as described in the

previous subsection. Furthermore, we derive three other indicators from hypotheses concerning the behaviour of falsifiers fabricating data. Schäfer *et al.* (2005) assume that falsifiers have a tendency to answer every question, thus producing less missing values. Furthermore, in line with Porras and English (2004), they expect falsifiers to choose less extreme answers to ordinal questions. Hood and Bushery (1997) hypothesize that falsifiers will "try to keep it simple and fabricate a minimum of falsified data" (Hood and Bushery 1997, page 820).

Based on these assumptions, we calculate three proportions, which serve as indicator variables in the multivariate analyses along with the χ^2 -value. The three indicator variables are calculated as follows:

- The 'item-non-response-ratio' is the proportion of questions which remain unanswered in all questions. We expect this ratio to be lower for falsifiers than for honest interviewers.
- The 'extreme-answers-ratio' refers to answers which are measured in ordinal scales. The ratio indicates the share of extreme answers (the lowest or highest category on the scale) in all ordinal answers. According to the above-mentioned assumptions, this ratio should also be lower for falsifiers.
- The 'others-ratio' refers to questions which, besides several framed responses offer the item 'others' as a possible answer. The choice of this item requires the explicit declaration of an alternative. If falsifiers tend to keep it simple, we can expect them to prefer the framed responses to the declaration of an alternative. Thus, this ratio too (calculated as the proportion of 'others' answers in all answers where the others item is selectable) should be lower for falsifiers.

Of course, the list of indicator variables, which might be included in the cluster analysis, can be extended. Generally, it is possible to derive many more of those variables from hypotheses on the behaviour of interviewers who fabricate data or to use those which have already been proposed in the literature, albeit not in the context of cluster analysis. For example, based on the assumption that falsifiers try to fabricate a minimum of falsified data, Hood and Bushery (1997) expect them to disproportionately often select the answer 'No' to questions, which either lead to a set of new questions or avoid it (assuming that 'No' is generally the answer that avoids further questions). So one could calculate the ratio of 'No' answers to such questions and use this ratio as a variable in the cluster analysis. We do not use this ratio, as two slightly different versions of the questionnaire were used in our empirical sample. There are only a small number of questions that lead to new questions or avoid

them depending on the answers, which are identical in both versions of the questionnaire.

Furthermore, when computer assisted interviewing allows the use of date and time stamps as discussed by Bushery *et al.* (1999), the average time needed to conduct an interview or the number of interviews conducted in one day might serve as indicators. Panel surveys offer some additional information to construct indicators. Stokes and Jones (1989) propose to compare the actual rate of non-matched household members in an interviewer's questionnaires to expected nonmatch rates that are calculated conditional on several household characteristics. The authors employ this procedure in the post-enumeration survey that is conducted as follow-up survey for the U.S. Census. If the actual rate of nonmatches strongly exceeds the expected rate, the authors consider this to be an indicator for fabricated data. Generally, this approach is applicable as soon as one has two or more waves of a panel survey available.

It becomes obvious that the first steps of our approach consist of examining the structure of the questionnaire and other types of data like date or time stamps collected during the survey process. Then one might consider which indicators could be derived from those sources that are likely to differ between falsifiers and honest interviewers. Another approach is the use of data mining techniques to identify patterns that are common in fabricated data or patterns in which fabricated data differs from honestly collected data (Murphy, Eyerman, McCue, Hottinger and Kennet 2005). If those patterns are detected, they might be used as indicators instead of deriving indicators from hypothesis on falsifier behaviour. However, this approach requires a huge dataset with known cases of falsification in order to conduct the data mining process. Such a dataset is not always available.

3. Results

3.1 Data sources

The data used in this study are derived from household surveys conducted in November 2007 and February 2008 in a Commonwealth of Independent States (CIS) (*i.e.*, former Soviet Union) country. The survey was part of an international research project on land reforms and rural poverty. We intended to interview 200 households in four villages in 2007. After identifying that all interviews had been fabricated in the first surveyed village we broke the survey off and started a new round with new interviewers in other villages in February 2008. All villages had been selected by qualitative criteria like the agricultural production structure and the implementation of land reforms. The households within one village had been selected by random sample based on household lists, which were provided by the

mayors of the villages. This procedure not only assured that all households had been selected at random, but also provided the basis for reinterviews as all households were exactly defined. However, these reinterviews were not planned in the very beginning. Because the households rarely owned telephones, check-calls were not possible and reinterviews in these households were associated with high costs and expenditure of time for traveling to the village for a face-to-face reinterview. Five interviewers were engaged in the first 2007 survey. Two of them had been the local partners of the research project. They had been involved in the development of the questionnaire and were responsible for the coordination of the surveys in their country. The other three interviewers were students hired by the partners. The questionnaire was composed of different sections with regard to household characteristics, resource endowment as well as income and expenditures. Most of the questions were closed questions. Only a few questions included a scale. Metric variables were collected for household expenditures like leasing or buying land, seeds, fertilizer or taxes and household income from different sources like agricultural or non-agricultural self employment and public or private transfers.

When the interviews of the 2007 survey were conducted, none of the German researchers were present in the villages. The questionnaires were collected right after the survey of the first village. In a first review of the questionnaires, we became suspicious because the paper of the questionnaires looked very clean and white. There was no dirt or dog-ears on the paper. Comparing the answers of different questionnaires of one interviewer we found two questionnaires with identical answers. Considering the fact that we asked for the amount of income from different sources in metric numbers it was very unlikely that the answers of two questionnaires would have been identical. Not getting any explanations from the project partners, we reinterviewed a sub-sample of 10% of the original sample face-to-face. None of the reinterviewed households reported having been surveyed. After detecting the fabrication of the interviews, the partners acknowledged that all interviews had been fabricated. As a matter of course, we stopped working with all interviewers and partners and implemented a new local research group.

In February 2008, the survey was repeated in the same country. As mentioned before, we selected new villages and households according to the above-mentioned criteria. We hired nine students for the interviews and arranged the survey with on-site supervision. In most cases, the interviews took place in a school or the city hall so that we could monitor all interviewers. When the interviews took place in the houses of the surveyed families we attended some of them. Due to this procedure, we presume that the questionnaires from the 2008 survey are not fabricated.

In this paper, we use a total of 250 household interviews by 13 interviewers, of which four were falsifiers from the 2007 survey (the interviews submitted by one falsifier were excluded as he filled in only three questionnaires) who definitely faked the results, referred to as F1-F4, and nine interviewers who are supposed to be honest, labelled H1-H9. Table 1 provides an overview of the number of questionnaires per interviewer, which were included in the analysis.

Table 1
Number of questionnaires per interviewer

Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Number of questionnaires	10	12	10	10	22	23	23	24	23	23	23	23	24

3.2 Cluster analysis

In this subsection, we present the results of the cluster analysis. Based on the results, we evaluate the success of our procedure in identifying interviewers who fabricate data. As already mentioned, we use four indicator variables in the cluster analysis: the item-non-response ratio, the proportion of extreme ordinally scaled answers in all ordinally scaled answers referred to as extreme ratio, the proportion of answers where the others item including an alternative was selected in all answers which offered this item (referred to as others ratio) and the χ^2 -value stemming from the comparison of the leading digit distribution in the questionnaires of an interviewer with the respective distribution in all other questionnaires.

Table 2 provides the values of the four indicator variables included in the cluster analysis for all 13 interviewers. It shows that the item-non-response ratio and the others ratio are clearly lower for the four falsifiers than for the honest interviewers. F1 and F4 have not chosen the others item at all. For the extreme ratio, things seem to be less clear. All the values range between 40% and 70% except the value of interviewer F1, which is clearly lower. The χ^2 -values are quite high for falsifiers F2 and F4. The values of the other two falsifiers do not differ much from the ones observed for honest interviewers.

The general idea of cluster analysis is to identify subgroups of elements in a space of elements that are all characterized by multivariate measurements (see Härdle and Simar (2007) for an introduction to cluster analysis). In the first step, a measure to determine either distance or similarity between elements has to be chosen. In the second step, elements are assigned to different subgroups or clusters. Elements within one cluster should be similar according to the selected measure whereas elements in different clusters should be distant. There is a large variety of methods according to which elements can be assigned to

clusters whereby the number of clusters might either be fixed or determined by the cluster method.

Table 2
Values of the variables included in the cluster analysis for each interviewer (all values except χ^2 -value in percent)

Interviewer	Item-Non-Response	Others	Extreme	χ^2 -value
F1	1.36	0.00	28.33	19.63
F2	0.71	0.65	40.85	29.70
F3	0.68	2.33	56.90	11.34
F4	0.51	0.00	58.62	27.33
H1	3.85	18.01	65.12	14.48
H2	1.99	2.40	59.42	6.91
H3	3.10	9.47	70.07	15.49
H4	4.52	13.04	56.43	16.61
H5	1.18	4.48	70.07	12.16
H6	3.46	1.37	50.75	15.42
H7	2.51	12.72	45.65	9.11
H8	1.77	10.95	69.85	3.63
H9	0.14	1.61	69.44	19.14

We measured distance as squared Euclidian distance and employed several cluster procedures in order to check the robustness of the results. In all cases, the interviewers have been clustered in two groups with the intention to obtain one 'falsifier group' and one 'honest interviewer group.' The advantage of this approach is that a clear classification is obtained. In contrast, when one of the indicator variables is examined separately, it is not clear where to draw the line separating falsifiers and honest interviewers. Before conducting the cluster analysis, we standardized all variables on a mean of zero and on a variance of unity. This eliminates the scale effect as distances are measured in standard deviations and not in different units.

The first clustering method we use is hierarchical clustering. This is a standard procedure that can also be applied to larger datasets and is implemented in standard statistical software packages. Hierarchical clustering merges clusters step by step, combining the two closest clusters. At the beginning, every element is considered as a separate cluster. We measure distance between two clusters as the average squared Euclidian distance between all possible pairs of elements with the first element of the pair coming from one cluster and the second element from the other cluster. We used the software package STATA with the option 'average linkage' to conduct the hierarchical cluster analysis.

In hierarchical cluster analysis, two elements will stay in the same cluster once they are merged together. Thus, the procedure does not necessarily lead to a global optimum with regard to a given distance measure. In our case the relatively low number of interviewers allows us to conduct an alternative analysis by simply examining all possible cluster compositions and select the best one with regard to a certain target function. (The analysis was carried out in MATLAB, the program code is available upon request.)

This procedure is clearly superior to hierarchical clustering as it ensures that the globally optimal cluster composition is identified. However, we also provide the results of hierarchical clustering as it is rather feasible compared to the computationally intensive approach of trying all possible compositions when the number of interviewers rises. Alternatively, one might resort to heuristic optimization techniques.

When examining all possible cluster compositions we use two target functions. The first one combines the ideas that a large distance between the two cluster centers is eligible as well as a small distance between the elements of a cluster and the cluster center. We look for the cluster composition, which maximizes the following expression:

$$\frac{\sum_{i=1}^4 (\bar{d}_{1i} - \bar{d}_{2i})^2}{\sum_{j=1}^{n_1} \sum_{i=1}^4 (d_{ij} - \bar{d}_{1i})^2 + \sum_{j=n_1+1}^{13} \sum_{i=1}^4 (d_{ij} - \bar{d}_{2i})^2} \tag{3}$$

The index *i* represents the four different indicator variables, \bar{d}_{ai} with *a* = 1, 2 is the mean of variable *i* in cluster *a*, *j* symbolizes the different elements (interviewers) in cluster 1 and cluster 2, *d_{ij}* is the value of variable *i* for element *j*, and *n₁* is the number of elements in cluster 1. Thus the numerator measures the distance between the two clusters, the denominator the distance within clusters and distance is measured in squared Euclidian form.

Alternatively, it could be interesting to see what optimal cluster composition results if instead of maximizing Equation (3) the average squared Euclidian distance between all possible pairs of elements within one cluster is minimized. In fact, this idea is very similar to the relevant target function in the hierarchical cluster procedures we presented before. Our second distance measure, which this time is to be minimized, is calculated as follows:

$$\frac{\sum_{j=1}^{n_1-1} \sum_{k=j+1}^{n_1} SED_{jk} + \sum_{j=n_1+1}^{13-1} \sum_{k=j+1}^{13-1} SED_{jk}}{(n_1(n_1-1))/2 + ((13-n_1)(13-n_1-1))/2} \tag{4}$$

SED_{jk} is the squared Euclidian distance between elements *j* and *k*, calculated as $SED_{jk} = \sum_{i=1}^4 (d_{ij} - d_{ik})^2$. The numerator is the sum of distances between all possible pairs of elements in the same cluster. By dividing this sum by the number of possible pairs, one obtains the average within cluster distance.

Table 3 reveals the results of the three cluster procedures. In the hierarchical analysis with linkage between groups, the three falsifiers F1, F2 and F4 form cluster 1, falsifier F3 and all honest interviewers form cluster 2. Thus, we are able to separate both groups of interviewers, except one falsifier. However, without knowing from the outset which

interviewers fabricated data and which were honest, one would have to decide which of the two clusters contains the at risk interviewers. This can be done by comparing the means of the indicator variables for each cluster displayed in Table 4. For the hierarchical procedure, means of the item-non-response ratio and the others ratio are clearly lower in cluster 1. The same is true for the mean of the extreme ratio, albeit the difference between the two clusters is less striking. Finally, a higher mean of the χ^2 -value can be observed for cluster 1. Given these results, one would – according to the above mentioned hypotheses on the behaviour of falsifiers – correctly identify cluster 1 to be the cluster containing the at risk interviewers. We also tried to improve the results of the hierarchical clustering procedure using the cluster means displayed in Table 4 as starting point for the K-means analysis. However, the application of K-means clustering did not lead to any changes in the cluster composition.

Table 3
Results of the three employed clustering procedures

Hierarchical clustering													
Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Cluster	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance between clusters divided by distance within clusters													
Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Cluster	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance between elements in one cluster													
Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Cluster	1	1	1	1	2	2	2	2	2	2	2	2	1

Table 4
Indicator variable means by cluster for the three cluster compositions

Item-Non-Response			Others		Extreme		χ^2 -value	
Hierarchical clustering								
Cluster	1	2	1	2	1	2	1	2
Mean	0.86	2.32	0.22	7.64	42.60	61.37	25.55	12.43
Distance between clusters divided by distance within clusters								
Cluster	1	2	1	2	1	2	1	2
Mean	0.86	2.32	0.22	7.64	42.60	61.37	25.55	12.43
Distance between elements in one cluster								
Cluster	1	2	1	2	1	2	1	2
Mean	0.68	2.80	0.92	9.06	50.83	60.92	21.43	11.73

The cluster composition that maximizes Equation (3) is identical to the one obtained using hierarchical clustering. Consequently, as can be seen from Table 4, the indicator means within the two clusters are identical as well.

The cluster composition minimizing Equation (4) is slightly different. Cluster 1 now contains all falsifiers and one honest interviewer. The means of the indicator variables again clearly indicate cluster 1 to be the cluster containing the at risk interviewers. This is a very satisfying result. All falsifiers are identified and only one false alarm is produced.

However, it should be kept in mind that this does not mean that this particular cluster method works best when applied to another dataset.

To evaluate to what extent a higher number of indicators leads to better results, we repeated our cluster approach based on Equations 3 and 4 with all possible combinations of indicators, including cases that only rely on one indicator. The results (see Table 7 in the appendix) generally indicate that an increasing number of indicators improves the results. However, there are also combinations with a smaller number of indicators that lead to similar results compared to those based on all four indicators. Determining which indicator composition is the best would require the highly subjective fixation of the relative cost caused by non-identified falsifiers compared to the cost caused by a false alarm. But one can determine which indicator compositions are not Pareto dominated in the sense that there is no other composition that exhibits less non-identified falsifiers (false alarms) and at the same time not more false alarms (non-identified falsifiers). The indicator composition including all four indicators is the only one that is not Pareto dominated no matter which equation is used. In contrast, compositions including only one indicator are Pareto dominated in six out of eight cases.

3.3 Discriminant analysis

Finally, we turn to the discriminant analysis to check whether the hypotheses on falsifiers' behaviour our cluster analysis is based upon are valid. Discriminant analysis can be used if the clusters are known in order to assess how well the indicators in the analysis can separate the different groups and whether group membership can be predicted correctly (see Härdle and Simar (2007) for an introduction to discriminant analysis). In a linear discriminant analysis, the coefficients b_0 and b_i of the discriminant function $D = b_0 + \sum_{i=1}^n b_i x_i$ are determined in such a way that they maximize a function that increases with the difference of the mean D -values of the two different groups and at the same time decreases with the differences of the D -values of elements within the groups. In our case, the x_i are our four indicator variables and we obtain two groups by separating falsifiers and honest interviewers.

We use prior probabilities corresponding to the relative group size (4/13 and 9/13) in order to predict group membership. Table 5 shows the results. Obviously the four variables allow a good separation of the falsifiers and the honest interviewers, as the group membership is correctly predicted in all cases but one.

As can be seen from Table 5 negative values of the discriminant function are associated with the falsifier group. Consequently, Table 6 indicates that three of the four coefficients' signs are in line with the expected falsifier

behaviour. Higher item-non-response and extreme ratios lead to a higher probability to observe an honest interviewer as does a lower χ^2 -value. The estimated coefficient for the others ratio is negative. Thus an increase in the others ratio *ceteris paribus* raises the probability that an interviewer has fabricated data. This might appear as a contradiction to our above-mentioned hypotheses. One possible explanation might be that the effect of the others ratio is already captured by the item-non-response ratio. In fact, the correlation coefficient between the two variables is quite high with a value of 0.71. The Wilks' lambda of the discriminant analysis is statistically significant on the 5%-level.

Table 5
Results of the discriminant analysis by interviewer

Interviewer	Predicted group	Actual group	Discriminant function
F1	1	1	-2.878
F2	1	1	-3.376
F3	2	1	-0.541
F4	1	1	-1.955
H1	2	2	1.828
H2	2	2	1.060
H3	2	2	1.747
H4	2	2	1.616
H5	2	2	0.706
H6	2	2	0.777
H7	2	2	-0.041
H8	2	2	1.765
H9	2	2	-0.710

Table 6
Standardized and non-standardized estimated coefficients (discriminant analysis)

Variable	Coefficient (non-standardized)	Coefficient (standardized)
Item-Non-Response	0.767	0.917
Others	-0.025	-0.129
Extreme	0.075	0.821
χ^2 -value	-0.092	-0.562
Constant	-4.250	-
Wilks' lambda (Prob > F)	0.0254	

4. Conclusion

Survey data are potentially affected by interviewers who fabricate data. Data fabrication is a non-negligible problem as it can cause severe biases. Even a small amount of fraudulent data might seriously impair the results of further empirical analysis. We extend previous approaches to identify at risk interviewers by combining several indicators derived directly from the survey data by means of cluster analysis. To demonstrate our approach, we apply it to a small dataset which was partially fabricated by falsifiers. The fact that we know the falsifiers from the outset allows us to evaluate the results of the cluster analysis and to furthermore conduct a discriminant analysis to reveal how well the two

groups of interviewers can be separated by the indicator variables. Different types of cluster analyses are conducted. All of them lead to the identification of an at risk interviewer cluster, with the item-non-response ratio and the others ratio being the clearest indicators. We are not able to identify falsifiers perfectly. However, in all cases the at risk interviewer contains a much higher share of falsifiers than the second cluster. The advantage of clustering is that one obtains a clear classification of interviewers who are at risk and the other interviewers, something that is not the case when indicators like the χ^2 -value are examined separately. Furthermore, it allows us to combine the information of several indicators. By investigating the performance of all possible subsets of indicators we find that generally a larger number of indicators is more apt to identify falsifiers. The fact that different clustering methods lead to different results should not necessarily be considered a shortcoming of our approach. Depending on how one weights the costs of an undetected falsifier relative to a false alarm, one might finally assign only those interviewers to the potential falsifier group that always fall into the at risk cluster, no matter what clustering method is applied (which would imply high costs of false alarms), one might assign all interviewers to the potential falsifier group that fall into the at risk cluster at least once (which would imply high costs of undetected falsifiers) or choose a solution in between.

The application to a small dataset demonstrates another merit of our approach: it was tested and worked well in a situation in which the number of questionnaires per interviewer was quite limited (three of the falsifiers only submitted 10 questionnaires). If a small number of questionnaires per interviewer is sufficient to perform the analysis, one might also think about implementing it during the main field period when interviewers have only submitted a certain

percentage of their questionnaires. Falsifiers could then be replaced by other interviewers who survey the units that should have been surveyed by the falsifiers.

Of course, when examining our results one has to keep in mind that we applied our method to a dataset in which a very severe form of data fabrication occurred: on the one hand we have falsifiers that faked all of their questionnaires (nearly) completely, on the other hand we have interviewers that (presumably) did all of their work honestly, which eases the discrimination between honest interviewers dishonest interviewers. Furthermore, with 13 interviewers, the size of our sample is quite limited. It would be interesting to explore the usefulness of our approach applied to larger datasets, given that the share of falsified interviews in large surveys has been found to be smaller than in our case. Additionally, larger datasets might allow the construction of additional indicators for the cluster analysis. If the survey has a reinterview program it would be possible to evaluate the usefulness of our approach by comparing the ‘success’ of a random reinterview with the success of a reinterview focusing on interviewers that were labeled as being at risk. We also intend to pursue the analysis in an experimental setting. An appropriate setting can ensure that one obtains a dataset which was partly collected by conducting real interviews and partly fabricated by telling some participants in the experiment to fill their questionnaires themselves.

Acknowledgements

Financial support of the Deutsche Forschungsgemeinschaft through the project ‘SPP 1292: Survey Methodology’ is gratefully acknowledged.

We furthermore thank John Bushery and four anonymous referees for providing useful comments on our paper.

Appendix

Table 7
Results of the cluster analyses based on Equations 3 and 4 for all possible cluster combinations

Item-Non-Response	Indicators			Equation 3		Equation 4	
	Others	Extreme	χ^2 -value	Undetected falsifiers	False Alarms	Undetected falsifiers	False Alarms
			X	2	0	1	1
		X		2	1	2	2
		X	X	2	0	1 ¹	0
	X			0 ¹	4	0	4
	X		X	2	0	0	2
	X	X		3	0	0	3
	X	X	X	1 ¹	0	1	1
X				0 ¹	4	0	4
X			X	2	1	0	2
X		X		3	0	- ²	-
X		X	X	1 ¹	0	1	1
X	X			0 ¹	4	0	4
X	X		X	1	1	0	2
X	X	X		0 ¹	4	0	4
X	X	X	X	1 ¹	0	0 ¹	1

¹ Indicator composition not Pareto dominated.
² Mean cluster values did not allow for an identification of the ‘at risk’ cluster.

References

- Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(1), 551-572.
- Biemer, P., and Stokes, S. (1989). The optimal design quality control sample to detect interviewer cheating. *Journal of Official Statistics*, 5(1), 23-29.
- Bushery, J., Reichert, J., Albright, K. and Rossiter, J. (1999). Using date and time stamps to detect interviewer falsification. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 316-320.
- Diekmann, A. (2002). Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung. Technical Report Manuskript 06/2002, Institut für Technikfolgenabschätzung (ITA), Wien.
- Donoho, S. (2004). Early detection of insider trading in option markets. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 420-429.
- Eyerman, J., Murphy, J., McCue, C., Hottinger, C. and Kennet, J. (2005). Interviewer falsification detection using data mining. In *Proceedings: Symposium 2005, Methodological Challenges for Future Information Needs*. Statistics Canada.
- Forsman, G., and Schreiner, I. (1991). The design and analysis of reinterview: An overview. In *Measurement Errors in Surveys*, (Eds., P.B. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman), New York: John Wiley & Sons, Inc, 279-301.
- Guterbock, T.M. (2008). Falsification. In *Encyclopedia of Survey Research Methods*, (Ed., P.J. Lavrakas), Sage Publications, Thousand Oaks, 1, 267-270.
- Härdle, W., and Simar, L. (2007). *Applied Multivariate Statistical Analysis*, 2nd Edition. Springer, Berlin.
- Hill, T. (1995). A statistical derivation of the significant digit law. *Statistical Science*, 10(4), 354-363.
- Hill, T. (1999). The difficulty of faking data. *Chance*, 26, 8-13.
- Hood, C., and Bushery, M. (1997). Getting more bang from the reinterviewer buck: Identifying 'At risk' interviewers. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 820-824.
- Li, J., Brick, J., Tran, B. and Singer, P. (2009). Using statistical models for sample design of a reinterview program. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 4681-4695.
- Murad, U., and Pinkas, G. (1999). Unsupervised Profiling for Identifying Superimposed Fraud. Lecture Notes in Computer Science, 1704, 251-261.
- Murphy, J., Baxter, R., Eyerman, J., Cunningham, D. and Kennet, J. (2004). A system for detecting interviewer falsification. Paper Presented at the American Association for Public Opinion Research 59th Annual Conference.
- Nigrini, M. (1996). A taxpayers compliance application of Benford's law. *Journal of the American Taxation Association*, 18, 72-91.
- Nigrini, M. (1999). I've got your Number. *Journal of Accountancy*, 187(5), 79-83.
- Porras, J., and English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 4223-4228.
- Saville, A. (2006). Using Benford's law to predict data error and fraud - An examination of companies listed on the JSE Securities Exchange. *South African Journal of Economic and Management Sciences*, 9(3), 341-354.
- Schäfer, C., Schräpler, J., Müller, K. and Wagner, G. (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. *Schmollers Jahrbuch*, 125, 183-193.
- Schnell, R. (1991). Der einfluss gefälschter Interviews auf survey ergebnisse. *Zeitschrift für Soziologie*, 20(1), 25-35.
- Schräpler, J., and Wagner, G. (2003). Identification, Characteristics and Impact of Faked Interviews in Surveys - An analysis by means of genuine fakes in the raw data of SOEP. IZA Discussion Paper Series, 969.
- Schreiner, I., Pennie, K. and Newbrough, J. (1988). Interviewer falsification in census bureau surveys. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 491-496.
- Scott, P., and Fasli, M. (2001). Benford's law: An empirical investigation and a novel explanation. CSM technical report, Department of Computer Science, University Essex.
- Stokes, L., and Jones, P. (1989). Evaluation of the interviewer quality control procedure for the post-enumeration survey. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 696-698.
- Swanson, D., Cho, M. and Eltinge, J. (2003). Detecting possibly fraudulent data or error-prone survey data using Benford's law. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 4172-4177.
- Thiprungsri, S. (2010). Cluster Analysis for Anomaly Detection in Accounting Data. Collected Papers of the Nineteenth Annual Strategic and Emerging Technologies Research Workshop San Francisco, California.
- Turner, C., Gribbe, J., Al-Tayyip, A. and Chromy, J. (2002). Falsification in Epidemiologic Surveys: Detection and Remediation (Prepublication Draft). Technical Papers on Health and Behavior Measurement. Washington DC: Research Triangle Institute. No. 53.

The application of graph theory to the development and testing of survey instruments

Steven Elliott¹

Abstract

This paper focuses on the application of graph theory to the development and testing of survey research instruments. A graph-theoretic approach offers several advantages over conventional approaches in the structure and features of a specifications system for research instruments, especially for large, computer-assisted instruments. One advantage is to verify the connectedness of all components and a second advantage is the ability to simulate an instrument. This approach also allows for the generation of measures to describe an instrument such as the number of routes and paths. The concept of a 'basis' is discussed in the context of software testing. A basis is the smallest set of paths within an instrument which covers all link-and-node pairings. These paths may be used as an economic and comprehensive set of test cases for instrument testing.

Key Words: Graph theory; Computer Assisted Interviewing (CAI); Questionnaire development; Software testing; Basis testing; Test cases.

1. Introduction

Graph theory is a branch of mathematics which deals with collections of nodes and links. A visual representation of a collection of nodes and links is referred to as a 'graph'. Graphs have been used in many areas of study to model real-world phenomena. The earliest examples appear in the analysis of transportation logistics (Berge 1976, page VII). In such analyses, a graph-theoretic approach is useful for determining such things as a maximally efficient set of paths to cover a number of locations. The locations are represented by the nodes of the graph, and the links represent routes from one location to another.

Graph theory has applications also in survey methodology. If the questions in a survey questionnaire are represented as nodes and the routes of flow between questions are represented as links, then a graph may be used to model a questionnaire. As such, many of the theorems and descriptive measures from graph theory pertain to questionnaires. In addition, the processes of documenting and testing survey instruments benefit from a graph-theoretic approach. For example, a documentation system that contains one table for questions and another for response alternatives has the ability to verify the connectedness of all instrument components as well as perform simulations of a working instrument. A testing procedure in which the set of test cases minimally spans the 'basis' of an instrument graph guarantees that all combinations of consecutive links and nodes are tested with the smallest possible number of cases.

A graph-theoretic representation is not necessary for the development, documentation, or testing of most survey instruments. In most cases, survey instruments have relatively few questions and the routing through an

instrument does not have many branching points. Examples of this are customer satisfaction surveys and short, paper-and-pencil surveys such as the U.S. Census. For these types of instruments, conventional documentation and testing procedures are adequate. However, large and complex surveys, like many current survey efforts, may benefit from a graph-theoretic approach. For example, the Canadian Financial Capability Survey (CFCS) is a survey that was conducted in 2009 to determine Canadians' knowledge and behavior with respect to financial decision making. It was a computer-assisted telephone interview comprised of 12 sections each of which had approximately 12 questions (Statistics Canada 2010). Another example is the Consumer Expenditure Surveys Quarterly Interview CAPI Survey (2010) conducted by the United States Department of Labor, Bureau of Labor Statistics. This survey has 22 sections most of which have 3 or more subsections, and within each subsection there may be as few as six or as many as 90 questions (US Bureau of Labor Statistics 2010). Either of these examples would be a good candidate for a graph-theoretic approach to documentation and testing.

This paper addresses the application of mathematical graph theory to survey research instruments. The next section of the paper which follows immediately below contains a description of a questionnaire as a graph and a delineation of the special properties that set apart a questionnaire graph from other types of graphs. The third section outlines the implications of a graph-theoretic representation on the structure of databases used for documentation/specifications systems for computer-assisted surveys. In Section 4, the specific features of graph-theoretic data structures are discussed. Sections 5 and 6 pertain to software testing and

1. Steven Elliott, Westat, Incorporated, 1500 Research Blvd., Rockville, MD 20850-3158, U.S.A. E-mail: sdelliott2@verizon.net.

the implications of graph theory on testing. A rationale is presented for the use of a 'basis' set of test cases which covers all pairs of linked nodes. This set of paths constitutes a comprehensive set of test cases for instrument testing.

2. A questionnaire as a graph

A graph may be represented as follows: $G = (V, E)$, where $V = \{v_1, v_2, v_3, \dots, v_n\}$ is a set of nodes or vertices and $E = \{(v_i, v_j), (v_i, v_k), \dots\}$ is a set of links or relations between pairs of vertices. Links are referred to as 'edges' in the terminology of graph theory, and hence the common usage of "E" to represent them (Chartrand 1985, page 27). A graph need not have any additional special characteristics. However, graphs which are attributed special characteristics are useful in modeling many phenomena in science and engineering. For example, graphs with undirected edges (*i.e.*, where both of the nodes attached to a link may be a predecessor or successor) may be used to model AC electric circuits, and graphs with directed edges may be used to model problems in traffic-pattern design. Other graphs with special characteristics are utilized to model networks in computer science, communications, sociology, and psychology.

In the case of survey questionnaires, the nodes of the graph represent different components or parts of a survey instrument. Most frequently, these are the substantive questions of a survey or decision points where routing is determined. The edges represent the response alternatives or outcomes associated with a node. Edges also represent the routing from one node to the next, and each edge has a unique predecessor and successor node. The graph depicted in Figure 1 represents a simple, 12-question survey instrument. The black circles (*i.e.*, nodes) represent the components of the instrument, and the lines connecting the black circles represent the edges that join one question to another. For example, the first node could represent a question with two response alternatives such as 'yes' and 'no'. The second node could represent a question with five response alternatives, where the first three alternatives branch to node 3, and the fourth and fifth alternatives branch to node 4.

When a graph is used to represent a questionnaire, there are a number of special properties that are attributed to the graph. These properties define the logical nature of a questionnaire. Bethlehem and Hundepool (2004) pointed out a number of these properties. First, a questionnaire has a starting node and an ending node. Second, all nodes other than the starting and ending nodes are connected. This means that for each node in the graph there is at least one route to it from the starting node, and one route away from it to the ending node. A third property of a questionnaire graph is that each of the edges is directed. This means that

the route of flow from one node to another is always in one direction. A fourth characteristic of a questionnaire graph is that it may have multiple edges between a single pair of nodes. Many types of graphs are restricted such that only one edge may join a pair of nodes. This restriction does not apply to a questionnaire graph, because questionnaires commonly have more than one response alternative leading from one question to another. A final characteristic is that looping structures are permitted. This means that a node may appear multiple times on a single route. Looping structures are used frequently in questionnaires to modify responses that are determined to be incorrect. For example, financial or time-usage questions may be checked with edits that loop back if component questions do not sum to the correct total.

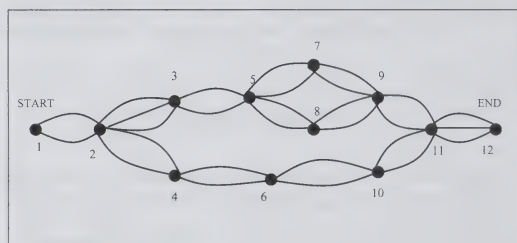


Figure 1 Representation of a Survey Instrument as a Graph

The characteristics of a questionnaire graph may be summarized as follows:

1. a starting node and an ending node,
2. connectedness (*i.e.*, each node is connected to the start and end nodes),
3. all edges are directed,
4. pairs of nodes may have multiple or parallel edges connecting them, and
5. nodes may appear more than once on a route.

Given a set of defining properties, it is possible to determine a number of descriptors including the number of routes and a basis. It is possible also to model a documentation system on the structure of the graph as illustrated in the next section.

3. Documentation and specification systems for survey questionnaires

Questionnaire documentation systems are typically one of two types: a text document or a relational database. For text-document systems, the information pertaining to a substantive question or other type of instrument component is most often presented as a section of the document. It consists of the question text, response alternatives, routing,

and instructions for programmers. The documentation system itself has no functionality aside from the search and print capabilities available in the word-processing software used to create the documentation. Systems using a relational database, on the other hand, are typically structured as a table where the rows represent the questions of the survey, and the columns represent attributes of the questions. Each record in the table is an n -tuple of question attributes. For example, the attributes of a question might include: name, sequence number, text of the question, response alternatives, routing information, and technical notes. One such specifications system is the *Tool for the Analysis and Documentation of Electronic Questionnaires* (TADEQ) (Bethlehem and Hundepool 2004). Other examples include systems developed at Westat Inc. for the *Medicare Current Beneficiary Survey* (MCBS) sponsored by the US Centers for Medicare and Medicaid Services (Medicare Current Beneficiary Survey: Overview 2010) and the *Medical Expenditure Panel Survey* (MEPS) sponsored by the US Department of Health and Human Services (MEPS: Survey Instruments and Associated Documentation 2010). These database systems have in common a structure of one primary table where each record represents a question.

Despite the advantages afforded by the straightforward nature of conventional systems, a specification system modeled like a graph has capabilities beyond those possible with a conventional structure. Before describing those capabilities and the necessary underlying structure, it should be noted that there are multiple ways in which a graph-theoretic data structure may be constructed (the interested reader is referred to Gibbons (1985, page 73) who described and categorized a number of those structures). The system proposed here is a relational list structure with two primary tables. One table represents the nodes of the graph, and the second table represents the edges. In the table representing nodes, each record or row represents an individual instrument component (*i.e.*, survey question, edit, or routing decision point). The second table represents edges where each record represents an individual edge (*i.e.*, a response alternative or a specific condition existing at a decision point). Each record from either table contains attributes associated with the record. Individual attributes are contained in the columns of the table. In the table of nodes, each column represents a specific attribute such as the component ID and component type. In the table of edges, each column represents an attribute such as the text of a response alternative. Two important distinctions between a documentation system with this structure versus a more conventional documentation system are: 1) the information pertaining to edges is not contained in the table for instrument components and 2) the table of edges (*i.e.*, links) contains identifiers for the predecessor and successor of an edge. As described in the

next section, these distinctions allow a documentation system to perform in ways not possible with conventional systems.

4. Features of a graph-based specifications system

The use of separate tables for nodes and links as the building blocks of a specifications systems has several advantages. Most important of these advantages is the ability to simulate an interview. A developer or tester can move through an instrument selecting response alternatives while being routed from one instrument component to another just as if they were administering the instrument to a respondent. Figure 2 is an example of a screen display for simulating an instrument. The component from which simulation begins is selected from this screen. Figure 3 is the actual simulation screen itself. It shows the current component with the question text or conditional in the center of the screen. The lower left is a display of all components from which one may have come in order to arrive at the current component (*i.e.*, predecessors). These are referred to as 'origination points' in the screen display. The lower right is a display of destination points or components to which one may go from the current component (*i.e.*, successors). Thus, one may move through an instrument one component at a time in either direction by selecting either an origination point or a destination point. In Figures 2 and 3, the questionnaire used as an example is one on general knowledge about cancer, and the question depicted in Figure 4 has only one predecessor and one successor. This will be the case for most survey questions, however if multiple predecessors or successors did exist, they would be listed in the display.

The ability to simulate the operation of a survey instrument is made possible because a separate table is utilized for links. This table may be queried to find all predecessors and successors for any component in the questionnaire. During the design phase of development, this feature can be used to insure that all sections and questions are properly connected and all routing is correct. In the testing phase of development, this feature may be used to perform side-by-side comparisons of an instrument and the specifications upon which it was built. A tester could have the specifications system simulating the instrument on one monitor while running the actual instrument on a second. Such comparisons can be used to check not only the wording and formatting of questions and response alternatives, but also to verify that the instrument is going to the appropriate question at the appropriate time. Reports of errors or problems may then be entered directly into the specification system as an attribute of an instrument component.

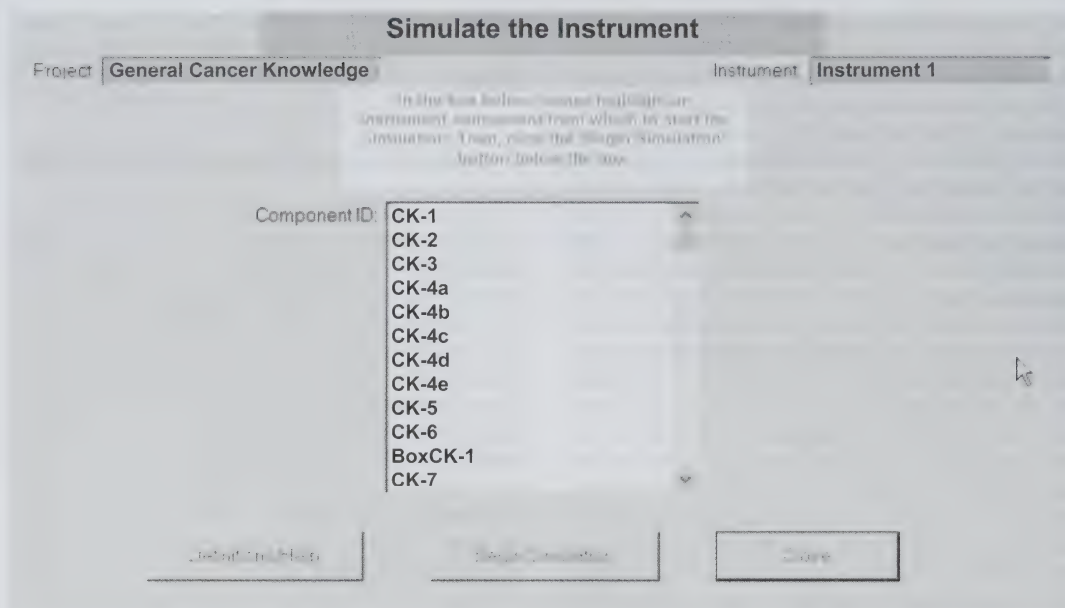


Figure 2 Begin simulation screen

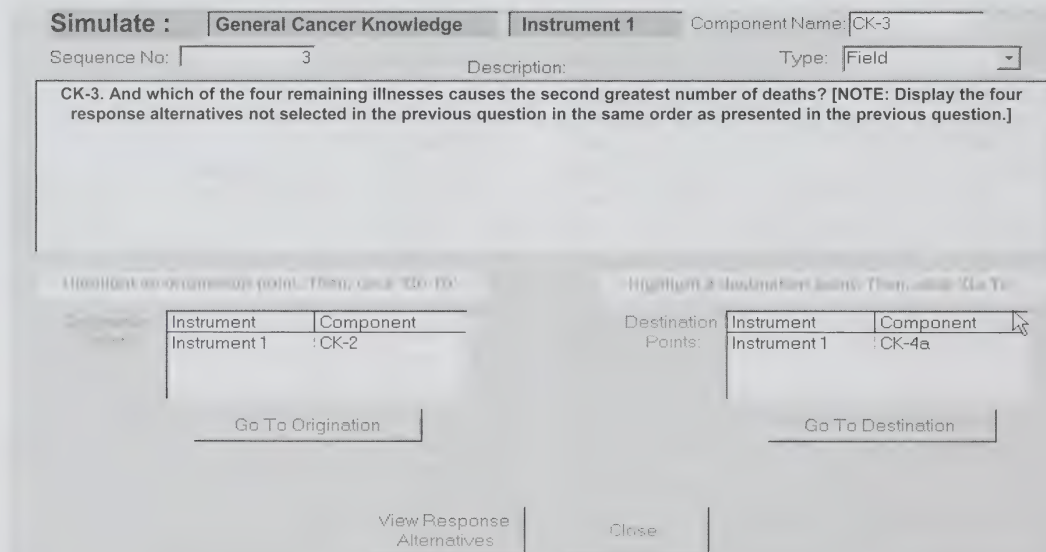


Figure 3 Simulation screen

Another method for evaluating the integrity of a questionnaire is to identify ‘orphan’ instrument components. Sometimes in the course of creating or modifying a questionnaire, an instrument component may become inaccessible. Such components are referred to as ‘orphans’. Since a table exists for links (*i.e.*, response alternatives and conditions), it is possible to run queries on this table to determine if a particular question appears as the successor to any link. If the question does not appear as a successor, then it is an orphan. Figure 4 contains the screen display for a listing of instrument components sorted by the frequency with which each appears as a successor. This is called an ‘Orphan Report’ in the figure. It shows that the first question in the survey has no origination points. This is as it should be since the first component cannot have predecessors. Any other component having zero origination points is an orphan. The orphan report is useful also in characterizing instrument components. For example, a question or component with a large number of originations may be the first question of a section devoted to handling premature terminations. Such a section is accessible from any other section of the interview, and therefore it would have a large number of predecessors.

5. Testing

Testing a computer-assisted survey instrument is the process of verifying that the behavior of the instrument is consistent with the design specifications. Several approaches have been utilized to accomplish this. One is to test first the building block components of a system, and

then move to increasingly larger and more integrated assemblages of components (*i.e.*, ‘bottom-up’ testing). Testing the building block components is referred to as ‘unit testing’ (Beizer 1995, page 5). After each of the building blocks has been tested separately, the blocks are assembled, and testing is concentrated on how the components interact. This is referred to as ‘integration testing’ (Hetzel 1984, page 11). The final stage of integration testing is ‘system testing’ where the entire system as a whole just as it would be used in a true production environment (Myers 1979, page 110).

Other approaches and terminology have also been applied to testing procedures. These include ‘black-box’, ‘white-box’, and ‘regression’ testing. In black-box testing, a program is treated as if it were in a black box where the inner workings not visible. Inputs and outputs are the only observable aspects of program function (Beizer 1995, page 8). White-box testing utilizes knowledge of the program code to decide how to conduct the tests and which cases are used in testing (Patton 2006, page 55). For example, a programmer might conduct a series of white-box tests such that every line of code is ‘exercised’ (*i.e.*, ‘code coverage’) or such that every branching point is exercised (*i.e.*, ‘branch coverage’). Regression testing is used to insure code integrity after changes or additions have been introduced to an operational program (Beizer 1995, page 235). Regression tests utilize a set of test cases. This set is selected such that each of the major branches of the program is exercised. Other types of testing (*e.g.*, alpha, beta, usability) are also used in software development, and there are many sources for a more comprehensive description of testing procedures (see Kaner, Falk and Nguyen 1999, page 277).

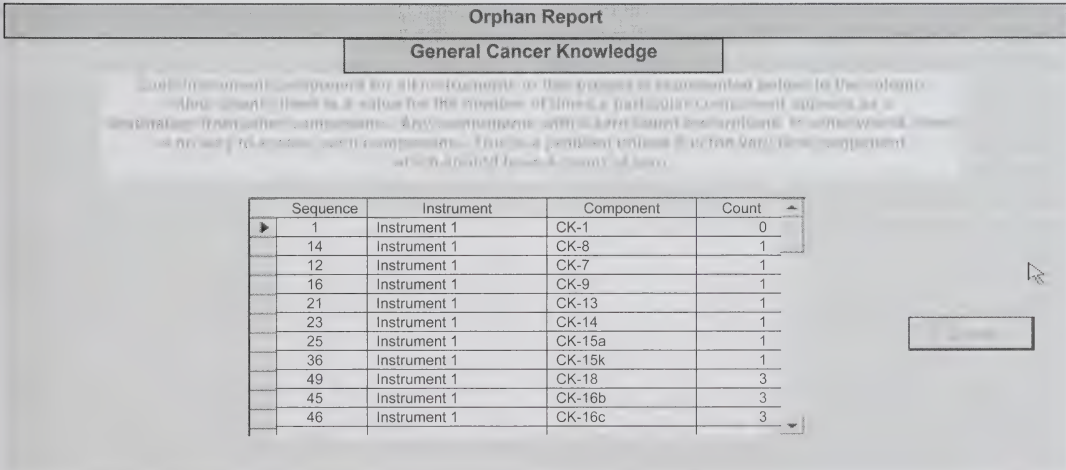


Figure 4 Orphan report

In any testing procedure, a major concern is testing bias. This results when some components or functionality of an instrument are excluded from testing. For example, questions which appear toward the end of a survey or in an obscure section may be more likely to be excluded. Testing bias is eliminated completely if a set of test cases is selected such that all instrument components, links between components, and aspects of functionality are included. However, given the length and complexity of some surveys, comprehensive testing is not a practical option. Consider, for example, the questionnaire represented in Figure 1. This questionnaire has only 12 questions and 28 response alternatives, and yet, there are 672 possible routes through the instrument. In large surveys such as those mentioned above, the number of routes could be well over 10,000. Thus, if comprehensive coverage is not a viable approach for large surveys, it is possible to avoid testing bias by taking a probability sample of potential test cases. A graph-theoretic approach can be useful in both the specification of the universe of test cases and in the determination of a rational approach to sampling test cases.

6. A graph-theoretic approach to testing

A universe of test elements can be defined in several different ways. One could use the elements already discussed - test cases, where each case is a mock interview. Alternatively, a universe of test elements could be survey questions, response alternatives, or any of a variety of combinations of questions and response alternatives. The discussion here is limited to test cases, and therefore, it will be helpful to provide precise definitions of a test case and two closely related terms, 'path' and 'route'.

A path is a unique, ordered set of nodes, which traverses an instrument from beginning to end. Each node in a given path, provided that it is not a starting or ending node, is linked to a predecessor and a successor (this definition is consistent with Bethlehem and Hundepool 2004). A unique path results whenever a component has more than one successor component. In Figure 1, multiple successors appear for components 2 and 4. These two branching nodes result in three paths:

Path 1 - 1, 2, 3, 5, 7, 9, 11, 12

Path 2 - 1, 2, 3, 5, 8, 9, 11, 12

Path 3 - 1, 2, 4, 6, 10, 11, 12

A 'route', on the other hand, is a unique, alternating series of nodes and links beginning with the starting node and terminating with the ending node. Like a path, a route must satisfy the properties of connectedness and direction. 'Route' is the graph-theoretic term which is synonymous

with what is commonly called a 'test case' in software testing. Since a route takes into account which link connects a pair of nodes, the number of routes in a graph is greater than or equal to the number of paths. The number of routes contained within a particular path is equal to the product of the number of links between each pair of nodes along the path. Thus for the example in Figure 1, the number of routes for each path is:

Path 1 - $2 \times 3 \times 2 \times 2 \times 2 \times 2 \times 3 = 288$

Path 2 - $2 \times 3 \times 2 \times 2 \times 2 \times 2 \times 3 = 288$

Path 3 - $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 3 = 96$

The total number of routes is the sum of routes over all paths (*i.e.*, $288 + 288 + 96 = 672$). A formula for computing the number of routes is:

$$\text{Routes} = \sum_i^P \prod_j^{NP_i} \text{links}_{ij}$$

where i represents the i^{th} path, P represents the total number of paths, j represents the j^{th} set of links on a given path, NP_i represents the number of pairs of connected nodes on a given path, and links_{ij} represents the number of links connecting a pair of nodes.

If a testing protocol is based on a sample of routes, then a minimum and comprehensive suite or universe of test cases is contained in the 'basis' of a graph. The term, 'basis', in this context is analogous to a 'basis' in geometry. The basis of a geometric space is a set of vectors which is sufficient to span the space, or in other words, a basis is a set of vectors sufficient to locate any point in the space. Likewise, the basis of a graph is a set of paths sufficient to include all predecessor-successor pairings of nodes. This implies that all nodes and at least one of the links between any connected pair of nodes are included. A basis is a subset of all possible paths. All questionnaires have a set of paths (P) in which each member satisfies the definition of a path as stated above (*i.e.*, a unique sequence of nodes). Within this set is a subset which has the special characteristic that each member path contains at least one pair of connected nodes that is not contained in any other path within the subset. This subset will be referred to as 'basis paths' (BP).

In order to gain a better understanding of the difference between the paths in BP and those in the complement of BP (*i.e.*, $P - BP$), consider the graph presented in Figure 5. The set of all paths (P) for the graph in Figure 5 is:

Path 1 - 1, 2, 4, 5, 7

Path 2 - 1, 2, 4, 6, 7

Path 3 - 1, 3, 4, 5, 7

Path 4 - 1, 3, 4, 6, 7

Any one of the four paths could be eliminated and the remaining three would include each pair of connected

nodes, and therefore any three constitutes a set of basis paths (BP). For example if Path 1 were eliminated, each of the node pairings would still be contained in Paths 2, 3, and 4. However, if both Paths 1 and 2 were eliminated, then node pairings 1 - 2 and 2 - 4 would be excluded. Thus, the set of two paths would be insufficient to span all of the independent sequences of nodes in the graph.

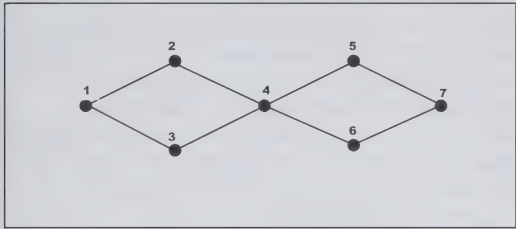


Figure 5 Representation of paths and basis paths

As illustrated above in Figure 1, many questionnaires encountered in practice have so many routes that testing all routes is not practical. Further, a typical route within an instrument has one or more similar routes which involve the same set of nodes, and these routes may be so similar that they differ by only a single, parallel link. Therefore, testing all routes would be not only impractical due to the large number of routes, but also redundant due to the similarity of many routes. The task for a test designer is to select a subset of routes that maximizes coverage and minimizes redundancy. This may be accomplished by using BP as a first step in sampling from the universe of routes. The utilization of BP in this manner is equivalent to beginning the sampling process with a purposive sample (Cochran 1977, page 10). Another way to think of this first step is as a redefinition of the universe of elements for the purpose of eliminating redundancy. This universe is comprehensive in its coverage, and it contains the smallest set of cases necessary to include all connected node pairs. A second stage of sampling could then be to select one or more routes from each of the paths contained in BP. This could be accomplished in several ways. One way would be to consider each path as a cluster of test cases and then take a probability sample from each cluster. Another way would be to select one route from each cluster by randomly selecting one parallel link at each node.

If one accepts the notion of basis testing, then it must be determined how much of the basis should be tested. If all paths in BP are tested, then the only elements of an instrument excluded from testing are redundant links. While redundant links may contain spelling or formatting errors, they are unlikely to contain routing errors. This stems from the nature of the programming task involved in creating

CAI instruments. Response alternatives are typically ‘bundled’ in the sense that alternatives which lead to the same next question are likely to be either all misdirected or none misdirected. For this reason, comprehensive testing of a basis is an effective method for minimizing errors of a type most likely to lead to loss of data.

On the other hand, non-comprehensive testing may be the only reasonable strategy if constraints due to time or level of effort exist and the number of paths in a basis is large. Despite the fact that any part of an instrument not tested may contain an error, any fraction of the paths in a basis may constitute an unbiased test. Thus, the percentage of paths to be included in a test should probably depend on factors specific to a particular development situation. For example, an instrument may contain modules which have been used previously or modules that have had only minor modification since previous use. These modules need not be tested as thoroughly as newer ones. As a general rule, a minimum sample of test cases should include each distinct section of an instrument in one or more paths, and paths should be included to cover all inter-sectional connections.

7. Discussion and conclusions

A graph-theoretic approach to software development has two major advantages over conventional approaches. First, it allows for a documentation system that can simulate the behavior of a computer-assisted interview. This is useful in verification of routing and as an aid to testers in side-by-side comparisons of instrument behavior versus design specifications. The second major advantage is in selecting cases for testing. The use of the basis of a questionnaire allows for the specification of a universe of test cases which covers all node pairings with a minimum number of paths. Probability sampling from this universe insures that no bias is incorporated into the testing procedures.

In practice, the first advantage can be achieved by structuring the database behind a specifications system such that it contains a table for nodes and a table for links. If the links table specifies a predecessor and a successor node, then queries of the tables will provide the functionality for verification of routing and simulation. The second advantage can be achieved with an algorithm for the identification of a basis. As pointed out by Poole (1995), one of the most important things to do when setting out to test software is to determine which test cases to use. He presented an algorithm for doing this that is based on the flowgraph of a program. Using a flowgraph for this purpose is useful as long as the program is not too large. With large and complicated programs, flow diagrams

become unwieldy. The same is true of large and complicated questionnaires (Bethlehem and Hundepool 2004). The appendix contains output from an algorithm which generates a basis, counts routes, and specifies basis paths for an example questionnaire graph (the algorithm used to generate the output appearing in the appendix is available from the author (sdelliott2@verizon.net). This algorithm does not handle looping structures as would be inherent in edits or 'go back' features. These structures may be tested as separate from the questionnaire graph. An algorithm which handles looping is under development).

A graph-theoretic approach is valuable also in that it allows for the use of a number descriptive measures of questionnaires such as the number of routes, the number of paths, cyclomatic complexity (cyclomatic complexity is a measure of complexity in software code (see Hetzel 1984; McCabe 1976; and Watson and McCabe 1996). It is equal also to the number of paths in the basis of a graph. For directed graphs where parallel links are not permitted, cyclomatic complexity (CC) = $L - N + 2$, where L is the number of links and N is the number of nodes), and several types of descriptive matrices (see appendix). Future enhancements to a graph-theoretic approach will likely involve such things as: 1) taxonomies for components, links, and errors; 2) secondary tables in the specification database containing attributes specific to different types of nodes and links; 3) sophisticated sampling plans for selecting test cases; and 4) purposive route sampling.

Taxonomies will promote the specification of special types of instrument components and the incorporation of secondary tables in the documentation system. An example of a special type of instrument component is one with a randomization feature. Such a component would be used in multi-phase respondent selection where a respondent reporting a particular disease, for example, has an increased probability of being routed to a follow-up section pertaining to that disease. In this case, the initial question pertaining to the disease may be a special type called 'respondent selection'. A secondary table in the documentation system for 'respondent selection' questions may have attributes pertaining to a random number generator such as generator seed and selection threshold.

Enhancements to sampling may include stratified sampling (Cochran 1977, page 89) and sampling with probability proportional to size (*i.e.*, PPS). Stratified sampling could be used to insure that all sections within a questionnaire are included with certainty. Paths would be stratified according to the sections they traverse. With PPS sampling, size might be a measure of path length, and the probability of selection

for a particular path would be dependent on the number of nodes included in the path. Thus, longer paths could be included with greater frequency. Purposive route sampling may be utilized for testing instrument characteristics other than programming errors. For example, later phases of questionnaire development might target specific sequences of questions for tests of the cognitive characteristics of an instrument.

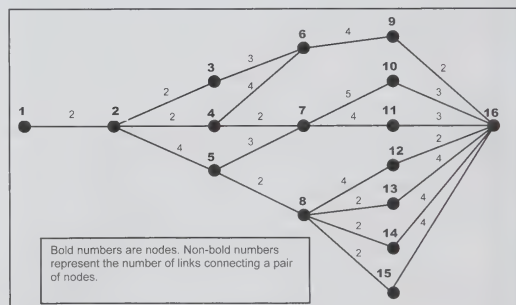
Other researchers in this area likely will provide further enhancements to the application of graph theory to questionnaire development. It does seem clear that graph theory lends itself well to the description, development, and testing of complex CAI instruments. The current trends in CAI usage seem to be in the direction of more sophisticated and larger instruments. For this reason, tools which help to document instrument components and identify errors are valuable to development efforts.

Acknowledgements

A significant portion of the work that went into this paper was supported by Westat, Incorporated Rockville Maryland. Substantial improvements and contributions to this paper were made by the reviewers and editors of *Survey Methodology*.

Appendix

Example of Basis Generation



Links (*i.e.*, excluding redundant links) = 23
Nodes = 16

Figure 6 Questionnaire graph

Table 1
Branches count for each node

Node Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Number of Branches	1	3	1	2	2	1	2	4	1	1	1	1	1	1	1	0

Table 2
Link matrix

Node	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		2														
2																
3																
4			2	2	4											
5						3										
6						4	2									
7							3	2								
8								2								
9									4							
10										5	4					
11												4	2	2	2	
12																
13																
14																
15																
16																

Each cell contains a value for the number of links between the row and column nodes.

Table 3
Path matrix

	1 st node	2 nd node	3 rd node	4 th node	5 th node	6 th node	7 th node	8 th node	9 th node	10 th node
Path 1	1	2	3	6	9	16				
Path 2	1	2	4	6	9	16				
Path 3	1	2	5	7	10	16				
Path 4	1	2	4	7	10	16				
Path 5	1	2	5	8	12	16				
Path 6	1	2	5	7	11	16				
Path 7	1	2	4	7	11	16				
Path 8	1	2	5	8	13	16				
Path 9	1	2	5	8	14	16				
Path 10	1	2	5	8	15	16				

Cell values represent nodes. Each row represents a path.
[Note: The paths in this example all have 6 nodes. However in general, all paths will not have the same number of nodes.]

Table 4
Link counts and number of routes for each path

	Node Pairings					Routes
	1 st to 2 nd	2 nd to 3 rd	3 rd to 4 th	4 th to 5 th	5 th to 6 th	
Path 1	2	2	3	4	2	96
Path 2	2	2	4	4	2	128
Path 3	2	4	3	5	3	360
Path 4	2	2	2	5	3	120
Path 5	2	4	2	4	2	128
Path 6	2	4	3	4	3	288
Path 7	2	2	2	4	3	96
Path 8	2	4	2	2	4	128
Path 9	2	4	2	2	4	128
Path 10	2	4	2	2	4	128

Paths = 10 Total Routes = 1,600
Cells represent the number of links between successive nodes in a path.

Table 5
Basis path matrix

	1 st node	2 nd node	3 rd node	4 th node	5 th node	6 th node	7 th node	8 th node	9 th node	10 th node
Basis Path 1	1	2	3	6	9	16				
Basis Path 2	1	2	4	6	9	16				
Basis Path 3	1	2	5	7	10	16				
Basis Path 4	1	2	4	7	10	16				
Basis Path 5	1	2	5	8	12	16				
Basis Path 6	1	2	5	7	11	16				
Basis Path 7	1	2	5	8	13	16				
Basis Path 8	1	2	5	8	14	16				
Basis Path 9	1	2	5	8	15	16				

Cell values represent nodes. Each row represents a basis path.

Table 6
Link counts and number of routes for each basis path

	Node Pairings					Routes
	1 st to 2 nd	2 nd to 3 rd	3 rd to 4 th	4 th to 5 th	5 th to 6 th	
Basis Path 1	2	2	3	4	2	96
Basis Path 2	2	2	4	4	2	128
Basis Path 3	2	4	3	5	3	360
Basis Path 4	2	2	2	5	3	120
Basis Path 5	2	4	2	4	2	128
Basis Path 6	2	4	3	4	3	288
Basis Path 7	2	4	2	2	4	128
Basis Path 8	2	4	2	2	4	128
Basis Path 9	2	4	2	2	4	128

Basis Paths = 9 Total Routes in Basis = 1,504

Cells represent the number of links between successive nodes from the Basis Paths Matrix above.

References

- Balakrishnan, V.K. (1997). *Graph theory*. New York: McGraw Hill, Inc.
- Beizer, B. (1995). *Black-box testing*. New York: John Wiley & Sons, Inc.
- Berge, C. (1976). *Graphs and hypergraphs*. New York: North-Holland Publishing Company - Amsterdam, London and American Elsevier Publishing Company, Inc.
- Bethlehem, J., and Hundepool, A. (2004). TADEQ: A tool for the documentation and analysis of electronic questionnaires. *Journal of Official Statistics*, 20, 233-264.
- Centers for Medicare and Medicaid Services (2010). *Medicare Current Beneficiary Survey: Overview*. August 2002. Internet address: https://www.cms.gov/LimitedDataSets/11_MCBS.asp.
- Chartrand, G. (1985). *Introductory Graph Theory*. New York: Dover Publications, Inc., Mineola.
- Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley & Sons, Inc.
- Cohen, J. (1997). *Design and methods of the Medical Expenditure Panel Survey Household Component*. Rockville (MD): Agency for Health Care Policy and Research. MEPS Methodology Report No. 1. AHCPR Pub. No. 97-0026.
- Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L. and O'Reilly, J.M. (1988). *Computer Assisted Information Collection*. New York: John Wiley & Sons, Inc.
- Gibbons, A. (1985). *Algorithmic Graph Theory*. Cambridge University Press.
- Harary, F., and Palmer, E. (1973). *Graphical Enumeration*. New York: Academic Press.
- Hetzel, W. (1983). *The Complete Guide to Software Testing, QED*. Massachusetts: Information Sciences, Inc., Wellesley.
- Kaner, C., Falk, J. and Nguyen, H. (1999). *Testing Computer Software*, 2nd Edition. New York: John Wiley & Sons, Inc.
- Medical Expenditure Panel Survey (MEPS) (2010). *Survey Instruments and Associated Documentation*. <http://www.meeps.ahrq.gov/meepsweb/>.
- Myers, G.J. (1979). *The Art of Software Testing*. New York: John Wiley & Sons, Inc.
- Patton, R. (2006). *Software Testing*, 2nd Edition. Sams Publishing, Inc.
- Poole, J. (1995). NISTIR 5737 – A Method to Determine a Basis Set of Paths to Perform Program Testing, National Institute of Standards and Technology, Gaithersburg, MD, November, 1995.

- Statistics Canada (2010). *Canadian Financial Capability Survey (CFC\$): Questionnaire 2009*. <http://www.statcan.gc.ca>.
- Statistics Netherlands (2002). *Blaise Developer's Guide*. Department of Statistical Informatics, Statistics Netherlands, Heerlen.
- US Bureau of Labor Statistics (2010). *Consumer Expenditure Surveys Quarterly Interview CAPI Survey 2010*. United States Department of Labor, <http://www.bls.gov/cex/capi/2010/cecapihome.htm>.
- Watson, A., and McCabe, T. (1996). NIST Special Publication 500-235 Structured Testing: A Testing Methodology Using the Cyclomatic Complexity Metric. National Institute of Standards and Technology, Gaithersburg, MD, September, 1996.

On sample allocation for efficient domain estimation

G. Hussain Choudhry, J.N.K. Rao and Michael A. Hidirolou¹

Abstract

Sample allocation issues are studied in the context of estimating sub-population (stratum or domain) means as well as the aggregate population mean under stratified simple random sampling. A non-linear programming method is used to obtain "optimal" sample allocation to strata that minimizes the total sample size subject to specified tolerances on the coefficient of variation of the estimators of strata means and the population mean. The resulting total sample size is then used to determine sample allocations for the methods of Costa, Satorra and Ventura (2004) based on compromise allocation and Longford (2006) based on specified "inferential priorities". In addition, we study sample allocation to strata when reliability requirements for domains, cutting across strata, are also specified. Performance of the three methods is studied using data from Statistics Canada's Monthly Retail Trade Survey (MRTS) of single establishments.

Key Words: Composite estimators; Compromise allocation; Direct estimators; Domain means; Non-linear programming.

1. Introduction

Stratified simple random sampling is widely used in business surveys and other establishment surveys employing list frames. The population mean $\bar{Y} = \sum_h W_h \bar{Y}_h$ is estimated by the weighted sample mean $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, where $W_h = N_h / N$ is the relative size of stratum h ($h = 1, \dots, L$) and \bar{Y}_h and \bar{y}_h are the stratum population mean and sample mean respectively. The well-known Neyman sample allocation to strata is optimal for estimating the population mean in the sense of minimizing the variance of \bar{y}_{st} subject to $\sum_h n_h = n$ where n is fixed or minimizing $\sum_h n_h$ subject to fixed variance of \bar{y}_{st} , where n_h is the stratum sample size. But the Neyman allocation may cause some strata to have large coefficients of variation (CV) of the means \bar{y}_h . On the other hand, equal sample allocation, $n_h = n/L$, is efficient for estimating strata means, but it may lead to a much larger CV of the estimator \bar{y}_{st} compared to that of Neyman allocation.

Bankier (1988) proposed a "power allocation" as a compromise between Neyman allocation and equal allocation. Letting $C_h = S_h / \bar{Y}_h$ be the stratum CV, the power allocation is

$$n_h^B = n \frac{C_h X_h^q}{\sum_h C_h X_h^q}, \quad h = 1, \dots, L \quad (1.1)$$

where X_h is some measure of size or importance of stratum h and q is a tuning constant. Power allocation (1.1) is obtained by minimizing $\sum_h \{X_h^q \text{CV}(\bar{y}_h)\}^2$ subject to $\sum_h n_h = n$, where $\text{CV}(\bar{y}_h)$ is the CV of the stratum sample mean \bar{y}_h . The choice $q = 1$ and $X_h = N_h \bar{Y}_h$ in (1.1) leads to Neyman allocation

$$n_h^N = n \frac{N_h S_h}{\sum_h N_h S_h}, \quad h = 1, \dots, L \quad (1.2)$$

and $q = 0$ gives equal allocation if $C_h = C$ for all h , where S_h^2 is the stratum variance. Bankier (1988) viewed values of q between 0 and 1 as providing compromise allocations. He gave a numerical example to illustrate how q may be chosen in practice. The choice $X_h = N_h$ and $q = 1/2$ in (1.1) gives "square root allocation" $n_h = n \sqrt{N_h} / \sum_h \sqrt{N_h}$ if $C_h = C$. Power allocation (1.1) and some other allocations generally depend on the variable of interest y and hence in practice a proxy variable with known population values is used in place of y .

Costa *et al.* (2004) proposed a compromise allocation based on a convex combination of proportional allocation, $n_h = n W_h$, and equal allocation $n_h = n/L$, see section 2.1. Longford (2006) made a systematic study of allocation in stratified simple random sampling by introducing "inferential priorities" P_h for the strata h and G for the population. In particular, he assumed that $P_h = N_h^q$ for a specified q ($0 \leq q \leq 2$), see section 2.4. He also studied the case of small strata sample sizes n_h in which case composite estimators of strata means \bar{Y}_h may be used.

The main purpose of our paper is to propose an "optimal" allocation method, based on non-linear programming (NLP), see section 2.3. It minimizes the total sample size $\sum_h n_h$ subject to specified tolerances on the CVs of the strata sample means \bar{y}_h and the estimated population mean \bar{y}_{st} . The case of indirect (composite) estimators of strata means is studied in Section 3. In Section 4, we study optimal sample allocation to strata when reliability requirements for domains, cutting across strata, are also specified.

1. G.H. Choudhry, Statistical Research and Innovation Division, Statistics Canada. E-mail: ghchoudhry@gmail.com; J.N.K. Rao, School of Mathematics and Statistics, Carleton University. E-mail: jrao@math.carleton.ca; M.A. Hidirolou, Statistical Research and Innovation Division, Statistics Canada. E-mail: mike.hidirolou@statcan.gc.ca.

The proposed method readily extends to multiple variables, but for simplicity we omit details. Using the optimal total sample size obtained from NLP, we make a numerical study of the performances of Costa *et al.* and Longford methods in terms of satisfying reliability requirements, Section 5.

2. Allocation for direct estimators

In this section, we consider direct estimators, \bar{y}_h , of strata population means, assuming stratified simple random sampling. The case of indirect estimators of strata means is studied in Section 3. Indirect strata estimators are used in the case of strata with small sample sizes n_h .

2.1 Costa *et al.* allocation

The sample allocation of Costa *et al.* (2004) is

$$n_h^C = k(nW_h) + (1-k)(n/L) \quad (2.1)$$

for a specified constant $k(0 \leq k \leq 1)$. This allocation reduces to equal allocation when $k = 0$ and to proportional allocation when $k = 1$. Formula (2.1) needs to be modified when $n/L > N_h$ for some h in a set of strata A . The modified allocation is

$$\tilde{n}_h^C = k(nW_h) + (1-k)n_h^0 \quad (2.2)$$

where $n_h^0 = N_h$ if $h \in A$ and $n_h^0 = (n - \sum_{h \in A} N_h) / (L - m)$ otherwise, where m is the number of strata in the set A . Note that when $k = 0$, (2.2) gives modified equal allocation. We study different choices of the constant k in the numerical study of Section 5, based on data from Statistics Canada's Monthly Retail Trade Survey (MRTS).

2.2 Longford allocation

Longford's (2006) method attempts to simultaneously control the reliability of the strata means \bar{y}_h and the estimated population mean \bar{y}_{st} by minimizing the objective function

$$\sum_{h=1}^L P_h V(\bar{y}_h) + (GP_+) V(\bar{y}_{st}) \quad (2.3)$$

with respect to the strata sample sizes n_h subject to $\sum_h n_h = n$, where $P_+ = \sum_h P_h$. The first component in (2.3) specifies relative importance, P_h , of each stratum h while the second component attaches relative importance to \bar{y}_{st} through the weight G . Longford (2006) assumed that $P_h = N_h^q$ for some constant $q(0 \leq q \leq 2)$. The term P_+ in (2.3) offsets the effect of the sizes P_h and the number of strata on the weight G .

Under stratified simple random sampling, the sample allocation minimizing (2.3) is

$$n_h^L = n \frac{S_h \sqrt{P_h'}}{\sum_h S_h \sqrt{P_h'}}, \quad h = 1, \dots, L \quad (2.4)$$

where $P_h' = P_h + GP_+ W_h^2$. If $q = 2$, then (2.4) does not depend on the value of G and it reduces to Neyman allocation, n_h^N , given by (1.2)

2.3 Nonlinear programming (NLP) allocation

We now turn to the NLP method of determining the strata sample sizes n_h subject to specified reliability requirements on both the strata sample means and the estimated population mean. Letting $\mathbf{f} = (f_1, \dots, f_L)^T$ with $f_h = n_h / N_h$, we minimize the total sample size

$$g(\mathbf{f}) = \sum_{h=1}^L f_h N_h \quad (2.5)$$

with respect to \mathbf{f} subject to

$$CV(\bar{y}_h) \leq CV_{0h}, \quad h = 1, \dots, L \quad (2.6)$$

$$CV(\bar{y}_{st}) \leq CV_0 \quad (2.7)$$

$$0 < f_h \leq 1, \quad h = 1, \dots, L \quad (2.8)$$

where CV_{0h} and CV_0 are specified tolerances on the CV of the stratum sample mean \bar{y}_h and the estimated population mean \bar{y}_{st} , respectively. Inequality signs are used in (2.6) and (2.7) because the resulting CVs for some strata h and/or for the aggregate may be smaller than the specified tolerances (Cochran 1977, page 122).

Letting $k_h = f_h^{-1}$, (2.5) becomes a separable convex function of the variables k_h ,

$$\tilde{g}(\mathbf{k}) = \sum_{h=1}^L N_h k_h^{-1} \quad (2.9)$$

We re-specify the constraints (2.6) and (2.7) in terms of relative variances so that the constraints are linear in the variables k_h . The relative variance (RV) of \bar{y}_h is the square of its CV,

$$RV(\bar{y}_h) = \frac{k_h - 1}{N_h} C_h^2 \quad (2.10)$$

Similarly, the relative variance of \bar{y}_{st} is the square of its CV,

$$RV(\bar{y}_{st}) = \bar{Y}^{-2} \sum_{h=1}^L W_h^2 \frac{k_h - 1}{N_h} S_h^2 \quad (2.11)$$

We used the SAS procedure NLP with the Newton-Raphson option to find the optimal k_h that would minimize (2.9) subject to

$$RV(\bar{y}_h) \leq RV_{0h}, \quad h = 1, \dots, L, \quad (2.12)$$

$$RV(\bar{y}_{st}) \leq RV_0, \quad (2.13)$$

$$k_h \geq 1, h = 1, \dots, L. \quad (2.14)$$

$RV(\bar{y}_h)$ and $RV(\bar{y}_{st})$ are given by (2.10) and (2.11) where $RV_{0h} = CV_{0h}^2$ and $RV_0 = CV_0^2$. By expressing the constraints as linear constraints and the objective function as a separable convex function, we achieve faster convergence of the re-formulated NLP. Denoting the solution to NLP as $\mathbf{k}^0 = (k_1^0, \dots, k_L^0)^T$, the corresponding vector of optimal strata sample sizes is given by $\mathbf{n}^0 = (n_1^0, \dots, n_L^0)^T$, where $n_h^0 = N_h / k_h^0$. We can modify (2.14) to ensure that $n_h^0 \geq 2$ for all h which permits unbiased variance estimation.

The NLP method can be readily extended to multiple variables y_1, \dots, y_P by specifying tolerances on the CVs of strata means and the estimated population mean for each variable ($p = 1, \dots, P$). If the number of variables P is not small, then the resulting optimal total sample size $n^0 = \sum_h n_h^0$ may increase significantly relative to n^0 for a single variable. Huddleston, Claypool and Hocking (1970), Bethel (1989) and others studied NLP for optimal sample allocation in the case of estimating population means of multiple variables under stratified random sampling.

3. Allocation for composite estimators

Longford (2006) studied composite estimators of strata means of the form

$$\hat{\theta}_h = \alpha_h \bar{y}_h^S + (1 - \alpha_h) \bar{y}_h \quad (3.1)$$

where \bar{y}_h^S is a synthetic estimator; here we take $\bar{y}_h^S = \bar{y}_{st}$. The MSE of $\hat{\theta}_h$ is

$$\begin{aligned} \text{MSE}(\hat{\theta}_h) &= V(\hat{\theta}_h) + [B(\hat{\theta}_h)]^2 \\ &= \alpha_h^2 \sum_{h=1}^L W_h^2 \frac{S_h^2}{n_h} + (1 - \alpha_h)^2 W_h^2 \frac{S_h^2}{n_h} \\ &\quad + 2\alpha_h(1 - \alpha_h) W_h \frac{S_h^2}{n_h} + \alpha_h^2 (\bar{Y}_h - \bar{Y})^2 \\ &\quad + \text{terms not depending on the } n_h. \end{aligned} \quad (3.2)$$

Longford (2006) showed that the optimal coefficient α_h in (3.1) minimizing (3.2) is approximately equal to $\alpha_h^* = S_h^2 / (S_h^2 + n_h \Delta_h^2)^{-1}$, where $\Delta_h = \bar{Y}_h - \bar{Y}$. He then replaced Δ_h^2 in α_h^* by its average over the strata, denoted by $\sigma_B^2 = L^{-1} \sum_h (\bar{Y}_h - \bar{Y})^2$, leading to $\alpha_h^* \approx (1 + n_h \omega_h^2)^{-1}$, where $\omega_h^2 = \sigma_B^2 / S_h^2$. The resulting MSE of $\hat{\theta}_h$ is approximated as

$$\text{MSE}(\hat{\theta}_h) \approx \frac{\sigma_B^2}{1 + n_h \omega_h}. \quad (3.3)$$

Longford's allocation is obtained by minimizing the objective function

$$\sum_{h=1}^L P_h \text{MSE}(\hat{\theta}_h) + (GP_+) V(\bar{y}_h) \quad (3.4)$$

with respect to the n_h . The resulting solution satisfies

$$\frac{P_h \sigma_B^2 \omega_h}{(1 + n_h \omega_h)^2} + (GP_+) W_h^2 \frac{S_h^2}{n_h} = \text{const.}, h = 1, \dots, L. \quad (3.5)$$

Longford used an iterative method to obtain the solution to (3.5) since it does not have a closed-form solution.

Our NLP procedure minimizes $g(\mathbf{f})$ given by (2.5) subject to

$$\text{RMSE}(\hat{\theta}_h) \leq \text{RMSE}_{0h}, h = 1, \dots, L; \quad RV(\bar{y}_{st}) \leq RV_0 \quad (3.6)$$

and (2.8), where $\text{RMSE}(\hat{\theta}_h) = \text{MSE}(\hat{\theta}_h) / \bar{Y}_h^2$ and RMSE_{0h} is a specified tolerance. The approximation (3) to $\text{MSE}(\hat{\theta}_h)$ is used in (3.6).

4. Allocation for domain estimation

Suppose that the population U is partitioned into domains ${}_dU$ ($d = 1, \dots, D$) that cut across the strata. Also, suppose that the estimators of domain means need to satisfy specified relative variance tolerances, ${}_dRV_0$, $d = 1, \dots, D$. We find the optimal additional strata sample sizes that are needed to satisfy the domain tolerances, using the NLP method.

An estimator of domain mean ${}_d\bar{Y} = {}_dN^{-1} \sum_{k \in {}_dU} y_k$ is the ratio estimator

$${}_d\hat{\bar{Y}} = \frac{\sum_{h=1}^L N_h n_h^{-1} \sum_{k \in s_h} {}_d\delta_k y_k}{\sum_{h=1}^L N_h n_h^{-1} \sum_{k \in s_h} {}_d\delta_k}, \quad (4.1)$$

where ${}_d\delta_k = 1$ if $k \in {}_dU$ and ${}_d\delta_k = 0$ otherwise, s_h is the sample from stratum h and ${}_dN$ is the size of domain d . The relative variance of the ratio estimator (4.1) is $\text{RV}({}_d\hat{\bar{Y}}) = V({}_d\hat{\bar{Y}}) / {}_d\bar{Y}^2$, where the variance $V({}_d\hat{\bar{Y}})$ is obtained by the usual linearization formula for a ratio estimator.

Let \tilde{n}_h denote the revised total sample size from stratum h so that the sample increase from stratum h is $\tilde{n}_h - n_h^0$. Let $\tilde{f}_h = \tilde{n}_h / N_h$ be the corresponding sampling fraction. We obtain the optimal $\tilde{\mathbf{n}} = (\tilde{n}_1, \dots, \tilde{n}_L)^T$ by minimizing the sample increase

$$g(\tilde{\mathbf{f}}) - \sum_{h=1}^L n_h^0 N_h = \sum_{h=1}^L (\tilde{f}_h - f_h^0) N_h \quad (4.2)$$

with respect to $\tilde{\mathbf{f}} = (\tilde{f}_1, \dots, \tilde{f}_L)^T$ subject to

$$f_h^0 \leq \tilde{f}_h \leq 1, \quad h = 1, \dots, L \quad (4.3)$$

$$RV({}_d\tilde{Y}) \leq {}_dRV_0, \quad d = 1, \dots, D. \quad (4.4)$$

As before, we reformulate the problem by expressing (4.2), (4.3) and (4.4) in terms of $\mathbf{k} = (\tilde{k}_1, \dots, \tilde{k}_L)^T$, where $\tilde{k}_h = \tilde{f}_h^{-1}$. This leads to minimization of the separable convex function

$$g^*(\mathbf{k}) = \sum_{h=1}^L N_h \tilde{k}_h^{-1} \quad (4.5)$$

with respect to $\tilde{\mathbf{k}}$ subject to the linear constraints

$$1 \leq \tilde{k}_h \leq k_h^0, \quad h = 1, \dots, L \quad (4.6)$$

and

$$RV({}_d\tilde{Y}) = {}_d\bar{Y}^{-2} \sum_{h=1}^L \left(\frac{N_h}{{}_dN} \right)^2 \frac{\tilde{k}_h - 1}{N_h} {}_dS_{eh}^2 \leq {}_dRV_0, \quad d = 1, \dots, D \quad (4.7)$$

where ${}_dRV_0$ is the specified tolerance, ${}_dS_{eh}^2$ denotes the stratum variance of the residuals ${}_de_k = {}_d\delta_k(y_k - {}_d\bar{Y})$ for $k \in U_h$ and U_h denotes the stratum population. Denote the resulting optimal \tilde{k}_h and \tilde{n}_h as \tilde{k}_h^0 and \tilde{n}_h^0 respectively, so that the optimal sample increase in stratum h is $\tilde{n}_h^0 - n_h^0$.

It can be shown that the minimization of total sample size subject to all the constraints $RV(\bar{y}_h) \leq RV_{0h}$, $h = 1, \dots, L$, $RV({}_d\bar{Y}) \leq {}_dRV_0$, $d = 1, \dots, D$, $RV(\bar{y}_{st}) \leq RV_0$, and $0 < \tilde{f}_h \leq 1$, $h = 1, \dots, L$ will lead to the same optimal solution, $\tilde{n}^0 = (\tilde{n}_1^0, \dots, \tilde{n}_L^0)^T$. However, domain reliability requirements may often be specified after determining \tilde{n}^0 .

5. Empirical results

In this section, we study the relative performance of different sample allocation methods, using data from the MRTS. Section 5.1 and 5.2 report our results for direct estimators and composite estimators of strata means, respectively. Results for the domain means are given in section 5.3.

5.1 Strata means: Direct estimators

For the empirical study, we used a subset of the MRTS population values restricted to single establishments. Strata sizes, N_h , strata population means, \bar{Y}_h , strata standard deviations, S_h , and strata CVs, $C_h = S_h/\bar{Y}_h$, are given in Table 1 for the ten provinces in Canada (treated as strata). For the NLP allocation, we have taken the CV tolerances as $CV_{0h} = 15\%$ for the strata means \bar{y}_h and $CV_0 = 6\%$ for the weighted sample mean \bar{y}_{st} , denoted Canada (CA).

The NLP allocation satisfying the specified CV tolerances resulted in a minimum overall sample size $n^0 = 3,446$. Table 2 reports the sample allocation n_h^0 and the

associated $CV(\bar{y}_h)$ and $CV(\bar{y}_{st})$ for the NLP allocation. It shows that the NLP allocation respects the specified tolerance $CV_0 = 6\%$, gives CVs smaller than the specified tolerance $CV_{0h} = 15\%$ for two of the larger provinces (QC: 11.4% and ON: 11.0%) and attains a 15% CV for the remaining provinces.

Table 1
Population values for the MRTS

Provinces	N_h	\bar{Y}_h	S_h	C_h
Newfoundland (NL)	909	963	1,943	2.02
Price-Edward-Island (PE)	280	712	1,375	1.93
New-Brunswick (NB)	1,333	1,368	3,200	2.34
Nova-Scotia (NS)	1,153	1,568	4,302	2.74
Quebec (QC)	11,135	2,006	4,729	2.36
Ontario (ON)	21,531	1,722	6,297	3.66
Manitoba (MN)	1,700	1,295	2,973	2.30
Saskatchewan (SK)	1,743	1,212	3,019	2.49
Alberta (AL)	5,292	1,698	5,358	3.16
British Columbia (BC)	7,803	1,291	4,013	3.11
Canada (CA)	52,879	1,654	-	-

Table 2
Equal, proportional, square root and NLP allocations and associated CVs (%)

Province	Equal		Proportional		Square-Root		NLP	
	n_h	CV_h	n_h	CV_h	n_h	CV_h	n_h	CV_h
NL	352	8.4	59	25.4	169	14.0	151	15.0
PE	280	0.0	18	44.1	94	16.2	104	15.0
NB	352	10.7	87	24.2	205	15.0	206	15.0
NS	352	12.2	75	30.6	191	18.1	259	15.0
QC	352	12.4	726	8.5	593	9.4	410	11.4
ON	352	19.3	1,403	9.4	824	12.5	1,056	11.0
MN	352	10.9	111	21.1	232	14.0	206	15.0
SK	352	11.9	114	22.6	234	15.2	238	15.0
AL	352	16.3	345	16.4	408	15.0	409	15.0
BC	352	16.2	508	13.3	496	13.5	407	15.0
CA	3,446	9.1	3,446	5.2	3,446	6.3	3,446	6.0

Using the optimal overall sample size 3,446, we calculated the sample allocations n_h and the associated $CV(\bar{y}_h)$ and $CV(\bar{y}_{st})$ for the modified equal allocation, proportional allocation and square-root allocation, reported in Table 2. It is clear from Table 2 that the modified equal allocation is not suitable in terms of satisfying specified CV tolerances because it leads to $CV(\bar{y}_{st}) = 9.1\%$ which is significantly larger than the specified $CV_0 = 6\%$. Also, under the modified equal allocation, $CV(\bar{y}_h)$ equals 19.3%, 16.3% and 16.2% for the larger provinces ON, AL and BC respectively. Note that for the smallest province PE Table 2 gives $CV(\bar{y}_h) = 0\%$ for the modified equal allocation because for PE it gives $n_h = N_h$.

Turning to proportional allocation, Table 2 reports $CV(\bar{y}_{st}) = 5.2\%$ but it leads to considerably larger strata CVs relative to the specified 15% for seven of the provinces, ranging from 16.4% to 44.1%. On the other hand, Table 2 shows that square-root allocation offers a reasonable compromise in terms of desired CV tolerances. We have $CV(\bar{y}_{st}) = 6.3\%$ and $CV(\bar{y}_h) \leq 15\%$ for seven of the provinces and the three provinces with CVs greater than 15% are SK with 15.2%, PE with 16.2% and NS with 18.1%.

Table 3 reports the results for the Costa *et al.* allocation (2.1) with $k = 0.25, 0.50$ and 0.75 , using $n = 3,446$ obtained from NLP. We observe from Table 2 that the choice $k = 0.25$, which assigns more weight to equal allocation, is not satisfactory for the estimation of the population (Canada) mean: $CV(\bar{y}_{st}) = 7.2\%$, but performs well for strata means, except AL with $CV(\bar{y}_h) = 16.3\%$. On the other hand, the choice $k = 0.75$, which assigns more weight to proportional allocation, performs poorly in estimating provincial means, with $CV(\bar{y}_h)$ ranging from 16.2% to 21.4% for seven of the provinces, although $CV(\bar{y}_{st})$ is smaller than the desired tolerance, 6%. The compromise choice $k = 0.50$ performs quite well, leading to $CV(\bar{y}_{st}) = 6.4\%$ and $CV(\bar{y}_h)$ around 15% or less except for two provinces (NS and AL) with CVs of 17.0% and 16.5% respectively. Performance of the Costa *et al.* method with $k = 0.50$ and square-root allocation are somewhat similar, and both allocations do not depend on the variable of interest, y , unlike the Longford and NLP allocations.

We next turn to Longford's allocation (2.4) which depends on q and G . Table 4 provides results for $q = 0, 0.5, 1.0, 1.5$ and $G = 0, 10, 100$, using $n = 3,446$ obtained from NLP. For $q = 2.0$, Longford's allocation does not depend on G and in fact it reduces to the Neyman allocation (1.2) which minimizes $CV(\bar{y}_{st})$ for fixed n but leads to highly inflated $CV(\bar{y}_h)$, ranging from 16% to 85% for seven provinces. We see from this table that $CV(\bar{y}_h)$,

for a given q , increases with G rapidly while $CV(\bar{y}_{st})$ decreases slowly as G increases and in fact is virtually a constant ($\approx 5.1\%$) for $G > 100$ (values not reported here). Also, $CV(\bar{y}_h)$ for a given G , increases rapidly as q increases while $CV(\bar{y}_{st})$ decreases. Langford's allocation, for $q \geq 0.5$ and/or $G \geq 10$, leads to significantly larger $CV(\bar{y}_h)$ than the specified tolerance $CV_{oh} = 15\%$ for several provinces, even though $CV(\bar{y}_{st})$ respects the specified tolerance of 6%. On the other hand, for $q = 0$ and $G = 0$, $CV(\bar{y}_h)$ is below the specified tolerance except for BC with 15.7%, but $CV(\bar{y}_{st}) = 7.3\%$ significantly exceeds the specified tolerance. For $q = 1.0$ and $q = 1.5$, $CV(\bar{y}_{st})$ stays below 6% when $G = 0$, but $CV(\bar{y}_h)$ exceeds 15% for six provinces, ranging from 17.7% to 34.0% for $q = 1.0$ and 22.0% to 54.6% for $q = 1.5$. On the whole, Table 4 suggests that no suitable combination of q and G can be found that ensures that all the specified reliability requirements are satisfied even approximately.

Table 3
Costa *et al.*'s allocation and associated CVs (%) for $k = 0.25, 0.50$ and 0.75

Province	$k = 0.25$		$k = 0.50$		$k = 0.75$	
	n_d	CV_d	n_d	CV_d	n_d	CV_d
NL	278	10.1	205	12.4	132	16.2
PE	214	6.4	149	10.8	83	17.8
NB	286	12.3	219	14.5	153	17.8
NS	282	14.2	213	17.0	144	21.4
QC	446	10.9	539	9.9	633	9.1
ON	615	14.5	878	12.1	1,140	10.5
MN	292	12.2	231	14.0	171	16.6
SK	292	13.3	733	15.2	174	17.9
AL	350	16.3	349	16.3	347	16.4
BC	391	15.3	430	14.6	469	13.9
CA	3,446	7.2	3,446	6.2	3,446	5.6

Table 4
CVs (%) for Longford's allocation with $q = 0, 0.5, 1.0$, and 1.5

Province	$q = 0$			$q = 0.5$			$q = 1.0$			$q = 1.5$		
	$G = 0$	$G = 10$	$G = 100$	$G = 0$	$G = 10$	$G = 100$	$G = 0$	$G = 10$	$G = 100$	$G = 0$	$G = 10$	$G = 100$
NL	13.5	19.3	29.7	17.2	23.0	33.4	22.7	29.0	38.3	30.4	36.2	40.6
PE	12.7	20.4	34.6	21.4	29.6	48.5	34.0	45.4	67.3	54.6	67.3	85.6
NB	12.0	17.1	25.0	14.5	19.4	26.8	18.3	23.1	29.0	23.5	27.6	30.3
NS	11.1	16.7	25.5	14.2	19.5	27.9	18.7	24.1	30.9	24.9	29.4	32.8
QC	11.0	9.8	9.1	9.9	9.4	9.0	9.2	9.0	8.9	8.9	8.9	8.8
ON	14.9	9.8	8.7	12.3	9.5	8.6	10.5	9.1	8.5	9.3	8.7	8.5
MN	12.7	17.6	24.3	14.7	19.1	25.2	17.7	21.9	26.5	22.0	25.4	27.5
SK	13.6	18.9	25.9	15.7	20.5	26.9	19.0	23.5	28.3	23.5	27.0	29.4
AL	13.5	15.7	16.1	13.3	15.2	15.9	13.6	15.2	15.9	14.6	15.5	15.9
BC	15.7	16.1	15.4	14.7	15.4	15.3	14.3	15.0	15.1	14.5	15.0	15.1
CA	7.3	5.5	5.1	6.2	5.3	5.1	5.5	5.2	5.1	5.2	5.1	5.1

5.2 Strata means: Composite estimators

We now report some results for the composite estimators, $\hat{\theta}_{h^*}$ of strata means. We obtained the optimal total sample size as $n = 3,368$ using NLP and the reliability requirements (3.6). This value is slightly smaller than the optimal $n^0 = 3,446$ for the direct estimators. For the Longford allocation, we used $n = 3,368$ and calculated the sample allocation and associated CVs of the composite estimators $\hat{\theta}_h$ and the weighted mean \bar{y}_{st} for specified q and G , constraining n_h to be at least two. For the MRTS data we have used, the first term of (3.5) is small relative to the second term. As a result, the sample allocation is flat across G - values for a given q which means that the CVs for the Longford allocation did not vary significantly with G .

Therefore, we have reported results in Table 5 only for $G = 0$ and $q = 0, 0.5, 1.0$ and 1.5 . We note from Table 5 that $CV(\hat{\theta}_h)$ decreases with q for the two largest provinces (QC and ON) because the sample shifts from the smaller provinces to these two provinces as q increases. Also, $CV(\hat{\theta}_h)$ initially decreases for AL and BC but it starts increasing when q is large because the sample starts shifting to QC and ON from these provinces as well. Further, $CV(\hat{\theta}_h)$ increases for all other provinces with q except for NS for which it starts decreasing for large q because of larger synthetic component and very negligible bias. In particular, $CV(\hat{\theta}_h)$ increases rapidly for NL and PE because of very large bias.

Table 5
CVs (%) for the composite estimators using Longford's allocation: $G = 0$ and $q = 0, 0.5, 1.0$ and 1.5

Province	$q = 0$	$q = 0.5$	$q = 1.0$	$q = 1.5$
NL	12.7	17.0	24.2	37.3
PE	12.4	23.8	46.0	112.2
NB	10.4	12.8	16.1	20.4
NS	9.4	11.9	14.5	11.7
QC	10.3	9.0	8.3	8.0
ON	13.9	11.1	9.3	8.2
MN	11.2	13.1	16.0	20.3
SK	12.4	14.6	17.9	23.2
AL	11.4	11.2	11.5	12.2
BC	14.4	13.3	12.9	13.1
CA	8.0	6.3	5.4	5.6

On the other hand, $CV(\bar{y}_{st})$ decreases initially with q but starts increasing when q is large because most of the sample gets allocated to QC and ON and very little sample is assigned to the smaller provinces. It appears from Table 5 that the Longford allocation performs reasonably well only for $q = 0$ and $G = 0$, giving $CV(\hat{\theta}_h)$ less than 15% for all provinces at the expense of $CV(\bar{y}_{st}) = 8.0\%$.

5.3 Domain means

Establishments on the Canadian Business Register are classified by industry using the North American Industry Classification System (NAICS). NAICS is principally a classification system for establishments and for the compilation of production statistics. The industry associated with each establishment on the Canadian Business Register is coded to six digits using the North American Industry Classification System. There are 67 six digit codes associated with the Retail sector. These six digit codes are regrouped into 19 trade groups (TG) for publication purposes.

We took the trade groups as domains that cut across provinces (strata). The trade group with the smallest number of establishments is TG 110 (beer, wine and liquor stores) with 307 establishments and the TG with the largest number of establishments is TG 100 (convenience and specialty food stores) with 7,752 establishments. Establishments were coded to all the 19 trade groups in all but one province: in PE, establishments were coded to only 16 trade groups.

We applied NLP based on (4.5), (4.6) and (4.7), and obtained the optimal total sample size increase to meet specified reliability requirements on the domain estimators \bar{y}_d . We found that no increase in the total sample size is needed if the tolerance on $CV(\bar{y}_d)$ is less than or equal to 30% for each domain. If the tolerance is reduced to 25%, then the optimal total sample size increase is 622 and the total sample size after the increase is 4,068. If the tolerance is further reduced to 20%, then the optimal total sample size increase is 2,100 and the total sample size after the increase is 5,546, which is considerably larger than the original 3,446. Note that as the total sample size is increased, CVs of strata means \bar{y}_h and the weighted sample mean \bar{y}_{st} decrease.

6. Summary and concluding remarks

We have proposed a non-linear programming (NLP) method of sample allocation to strata under stratified random sampling. It minimizes the total sample size subject to specified tolerances on the coefficient of variation of estimators of strata means and the population mean. We considered both direct estimators of strata means and composite estimators of strata means. The case of domains cutting across strata is also studied. Difficulties with alternative methods in satisfying specified reliability requirements are demonstrated using data from the Statistics Canada Monthly Retail Trade Survey of single establishments. We also noted that NLP can be readily extended to handle reliability requirements for multiple variables. Compromise allocations that perform reasonably well in terms of reliability requirements are also noted.

Acknowledgements

The authors thank two referees and an associate editor for constructive comments and suggestions.

References

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 47-57.

Bankier, M. (1988). Power allocation: Determining sample sizes for sub-national areas. *The American Statistician*, 42, 174-177.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. New York : John Wiley & Sons, Inc.

Costa, A., Satorra, A. and Ventura, E. (2004). Using composite estimators to improve both domain and total area estimation. *SORT*, 28, 69-86.

Huddleston, H.F., Claypool, P.L. and Hocking, R.R. (1970). Optimum allocation to strata using convex programming. *Applied Statistics*, 19, 273-278.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 87-96.

North American Industry Classification System, Version 1.4 (2008). Catalogue 12F0074XCB. Statistics Canada.

Calibration alternatives to poststratification for doubly classified data

Ted Chang ¹

Abstract

We consider alternatives to poststratification for doubly classified data in which at least one of the two-way cells is too small to allow the poststratification based upon this double classification. In our study data set, the expected count in the smallest cell is 0.36. One approach is simply to collapse cells. This is likely, however, to destroy the double classification structure. Our alternative approaches allows one to maintain the original double classification of the data. The approaches are based upon the calibration study by Chang and Kott (2008). We choose weight adjustments dependent upon the marginal classifications (but not full cross classification) to minimize an objective function of the differences between the population counts of the two way cells and their sample estimates. In the terminology of Chang and Kott (2008), if the row and column classifications have I and J cells respectively, this results in IJ benchmark variables and $I+J-1$ model variables. We study the performance of these estimators by constructing simulation simple random samples from the 2005 Quarterly Census of Employment and Wages which is maintained by the Bureau of Labor Statistics. We use the double classification of state and industry group. In our study, the calibration approaches introduced an asymptotically trivial bias, but reduced the MSE, compared to the unbiased estimator, by as much as 20% for a small sample.

Key Words: Calibration; Poststratification; Prediction model.

1. Introduction

Suppose we have a population \mathcal{U} which is doubly stratified by two categorical variables whose indices are denoted (i, j) , $i = 1, \dots, I$, $j = 1, \dots, J$ and write \mathcal{U}_{ij} for the (i, j) -stratum. If a simple random sample \mathcal{S} of size n is taken and if y denotes the variable of interest a natural estimator for the total $T_y = \sum_{k \in \mathcal{U}} y_k$ is the poststratified estimator

$$\hat{t}_{yPS} = \sum_{i,j} N_{ij} \bar{y}_{ij} \tag{1}$$

where N_{ij} is the size of \mathcal{U}_{ij} and \bar{y}_{ij} is the sample mean of y over $\mathcal{S} \cap \mathcal{U}_{ij}$. This estimator is widely used as long as all the sample sizes n_{ij} of $\mathcal{S} \cap \mathcal{U}_{ij}$ are reasonably large.

What to do if some of the n_{ij} are small, or even zero?

The standard approach would be to collapse some of the cells until all the n_{ij} are big enough. However such a collapsing might not be possible in a way that maintains the double classification scheme: that is the indices j might depend upon i .

The poststratified estimator \hat{t}_{yPS} is a special case of a calibration estimator. Define for each $k \in \mathcal{U}$ the $I \times J$ vector variable $\mathbf{x}_k = (x_{1k}, \dots, x_{Jk})^T$ where $x_{jk} = 1$ if $k \in \mathcal{U}_{ij}$ and $x_{jk} = 0$ otherwise. The population total $T_{\mathbf{x}}$ of \mathbf{x} is $(N_{11}, \dots, N_{IJ})^T$ and letting $d_k = N/n$ be the sampling weight and $\beta = (N^{-1}n_{11}^{-1}N_{11}n, \dots, N^{-1}n_{IJ}^{-1}N_{IJ}n)^T$

$$\hat{t}_{yPS} = \sum_{k \in \mathcal{S}} d_k (\mathbf{x}_k^T \beta) y_k$$

$$T_{\mathbf{x}} = \sum_{k \in \mathcal{S}} d_k (\mathbf{x}_k^T \beta) \mathbf{x}_k.$$

These equations establish that if the benchmark variables \mathbf{x} are used, then \hat{t}_{yPS} is the resulting calibrated estimator of T_y .

Chang and Kott (2008) derived the asymptotic properties of a calibrated estimate of the form

$$\hat{t}_{y,zfV} = \sum_{k \in \mathcal{S}} d_k f(\mathbf{z}_k^T \hat{\beta}) y_k \tag{2}$$

where $\hat{\beta}$ minimizes an objective function of the form

$$Q(\beta) = \left(T_{\mathbf{x}} - \sum_{k \in \mathcal{S}} d_k f(\mathbf{z}_k^T \beta) \mathbf{x}_k \right)^T \mathbf{V}^{-1} \left(T_{\mathbf{x}} - \sum_{k \in \mathcal{S}} d_k f(\mathbf{z}_k^T \beta) \mathbf{x}_k \right). \tag{3}$$

In equations (2) and (3), \mathbf{z} is a vector of model variables whose length Q is at most the length P of the benchmark variables \mathbf{x} , f is a positive real valued function which Chang and Kott (2008) calls the back link function, and \mathbf{V} is some positive definite symmetric $P \times P$ matrix. \mathbf{V} is allowed to depend upon β as would occur if $\mathbf{V}(\beta)$ is some measurement of the variability of $\sum_{k \in \mathcal{S}} d_k f(\mathbf{z}_k^T \beta) \mathbf{x}_k$.

In Chang and Kott (2008), the realized sample \mathcal{S} is the respondents from an original sample with sampling weights d_k . The respondent sample \mathcal{S} is assumed to be a Poisson subsample of the original sample with Poisson probabilities $f(\mathbf{z}_k^T \beta_0)^{-1}$, for some β_0 . The asymptotic formulas derived there were under an asymptotic framework for this quasi-randomization (design based) model. We use the term quasi-randomization to remind ourselves that the assumed Poisson response mechanism is actually model based.

It should be noted that the use of calibration to correct for nonresponse goes back to Fuller, Loughin and Baker (1994), at least when $\mathbf{z} = \mathbf{x}$ and $f(\eta) = 1 + \eta$.

1. Ted Chang, University of Virginia, Department of Statistics, Charlottesville, VA, U.S.A. E-mail: tcc8v@virginia.edu.

We propose to use the Chang and Kott (2008) methodology with \mathbf{x} remaining as indicator variables for the complete $I \times J$ cross classification but letting \mathbf{z} be a vector of $I + J - 1$ indicator variables for the marginal classifications. In other words, we propose to rebalance the sample to come as close as possible, in the sense of minimizing (3), to the correct cell proportions in the complete cross classification, but requiring the rebalancing weights to depend only upon the marginal classifications.

The Chang and Kott (2008) framework applies in the presence of nonresponse (and/or noncoverage) if $f(\mathbf{z}_0^T \beta_0)^{-1}$ is the response (or combined response and coverage) probability. We note that poststratification, a special case of calibration, is often used for the purpose of nonresponse/noncoverage correction. In our test example below, there is no nonresponse or noncoverage to correct for, and hence, the Chang and Kott (2008) framework applies with $\beta_0 = \mathbf{0}$ for any f with $f(0) = 1$. In other words, if the calibration is used solely for the purposes of sample rebalancing, we can use Chang and Kott (2008) with almost any f . But if we are trying to correct for nonresponse and/or noncoverage, stronger assumptions are required.

It should be noted that raking is simply the calibrated estimate using the $I + J - 1$ indicator variables of the marginal classifications as both benchmark and model variables and using $f(\eta) = e^\eta$. Thus we will also explore the use of this back link function.

Section 2 gives the precise formulas for the estimators we will use in this study. Chang and Kott (2008) can be applied to derive sample based variance estimators and these derivations are given in the Appendix.

In Section 3, we give the results on an empirical study using the 2005 first quarter Quarterly Census of Employment and Wages, collected by the Bureau of Labor Statistics. We will restrict ourselves to the five states which we will denote by A, B, C, D, E and to five industry groupings denoted by 1, 2, 3, 4, 5. We will not further identify either the states or the industry groupings to prevent identification of the outlier in the discussion below. This population has 283,725 firms. From this population we will take Monte Carlo simple random samples of size $n = 200, 1,000, 5,000$ and use the double classification of state and industry group.

It should be noted that 0.18% of the population has the double classification of state E and industry grouping 5. Thus when $n = 200$, the expected sample size in this cell is 0.36 and poststratification using the double classification is out of the question.

Kott and Chang (2010) derives the properties of $\hat{t}_{y,zfV}$ using a model based framework. The models considered there do not apply with our selection of \mathbf{x} and \mathbf{z} variables. However, motivated by their approach, we examine in Section 4 the behavior of the estimator $\hat{t}_{y,zfV}$ defined by

equation (2), under highly simplified assumptions, including that $f(\eta) = 1 + \eta$. This leads in Section 5 to the choice of a new weight matrix \mathbf{V}^{-1} for use in (3). We then continue with our empirical exploration using this new estimator.

2. Mathematical formulas

In this section we list the formulas used in this study. They are all special cases of formulas in Chang and Kott (2008). We assume that a simple random sample of size n is taken from a population of size N and we use \mathcal{S} and r to denote the respondents from that sample and the size of \mathcal{S} . We assume that the calibration weight function has a β_0 such that $f(\mathbf{z}^T \beta_0)^{-1}$ is the response probability for an element with model variables \mathbf{z} . In particular, and without loss of generality, if there is no nonresponse problem, we assume $f(0) = 1$.

The same formulas work with noncoverage, in which case $f(\mathbf{z}^T \beta_0)^{-1}$ is the combined response/coverage probability.

We denote N_{ij} , \mathcal{S}_{ij} , and r_{ij} to be the population size, respondent sample, and respondent sample size in classification (i, j) . Although N_{ij} is assumed known, our methodology does not require the knowledge of the row and column classifications of nonrespondents.

We define $N_{i\cdot} = \sum_j N_{ij}$ and analogously define $N_{\cdot j}$.

We will use estimators for a total T_y of the form

$$\hat{t}_y = \frac{N}{n} \sum_i \sum_j \sum_{k \in \mathcal{S}_{ij}} w_{ij} y_k \quad (4)$$

where the adjustment weights w_{ij} are defined as below. These are all special cases of equations (2) and (3) when we use $\mathbf{V} = \mathbf{I}$.

The *calibrated margins* estimator uses $f(\eta) = 1 + \eta$ and defines $\mathbf{x} = \mathbf{z}$ to be $I + J - 1$ independent indicator variables for the marginal categories. In this case T_x is a vector of $N_{i\cdot}$ and $N_{\cdot j}$. The adjustment weights $f(\mathbf{z}_k^T \beta)$ have the form $w_{ij} = 1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}$ when \mathbf{z} is the vector of indicator variables for membership in the i^{th} and j^{th} row and column classifications respectively. Since the number of equations (the dimension of \mathbf{x}) equals the number of unknowns (the dimension of $\hat{\beta}$), we expect to be able to solve the equations

$$T_x = \sum_{k \in \mathcal{S}} d_k f(\mathbf{z}_k^T \beta) \mathbf{x}_k \quad (5)$$

exactly. Thus $\hat{\beta}_{i\cdot}, \hat{\beta}_{\cdot j}$ solve the linear equations of rank $I + J - 1$

$$N_{i\cdot} = \frac{N}{n} \sum_j (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}$$

$$N_{\cdot j} = \frac{N}{n} \sum_i (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij},$$

which easily follows from (5).

The *calibrated cell counts* estimator uses $f(\eta) = 1 + \eta$ and defines \mathbf{x} to be the IJ indicator variables for the complete cross classification and \mathbf{z} to be $I + J - 1$ independent indicator variables for the marginal categories. In this case $T_{\mathbf{x}}$ is a vector of N_{ij} and, since $\mathbf{V} = \mathbf{I}$, the adjustment weights $w_{ij} = 1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}$ minimize the objective function

$$\sum_i \sum_j \left[N_{ij} - \frac{N}{n} \sum_i \sum_j (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij} \right]^2.$$

The *raking* estimator uses $f(\eta) = e^\eta$ and defines $\mathbf{x} = \mathbf{z}$ to be $I + J - 1$ independent indicator variables for the marginal categories. Its adjustment weights are $w_{ij} = \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})$ where $\hat{\beta}_{i\cdot}, \hat{\beta}_{\cdot j}$ solve the $I + J$ equations

$$N_{i\cdot} = \frac{N}{n} \sum_j \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}$$
$$N_{\cdot j} = \frac{N}{n} \sum_i \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}.$$

Since $\sum_i N_{i\cdot} = N = \sum_j N_{\cdot j}$, these $I + J$ equations yield only $I + J - 1$ constraints. It should be noted, however, that if a constant c is added to each $\hat{\beta}_{i\cdot}$ and subtracted from each $\hat{\beta}_{\cdot j}$, the w_{ij} are not changed.

The *exponential calibrated cell counts* estimator uses $f(\eta) = e^\eta$ and defines \mathbf{x} to be the IJ indicator variables for the complete cross classification and \mathbf{z} to be $I + J - 1$ independent indicator variables for the marginal categories. Its adjustment weights $w_{ij} = \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})$ minimize the objective function

$$\sum_i \sum_j \left[N_{ij} - \frac{N}{n} \sum_i \sum_j \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij} \right]^2.$$

Chang and Kott (2008) give formulas for sample based estimation of the variance of \hat{t}_y . In the appendix, we apply these formulas to the four estimators above.

3. Empirical study

The population we use here is the data from the 2005 first quarter Quarterly Census of Employment and Wages (QCEW), restricted to five states and five industry groupings. The QCEW is compiled from mandatory reports to state employment offices and hence is virtually a census and the data we used is the complete QCEW for these five states and five industry groupings. This population has $N = 283,725$ firms, divided as in Table 1.

The response variables y are total employment and total (quarterly) wages. For these variables $T_y = 2,981,364$ for total employment and $T_y = 2,334,400$ (in tens of thousands of dollars) for total wages. In this study, we took 10,000 samples of sizes $n = 200, 1,000, 5,000$. For each of the 4 estimators, we report the estimated bias, standard error, and root mean square error. We also report square root of the mean of the estimated variances using the first term of equation (15). For purposes of comparison, we report the theoretical and empirical values for the unweighted estimator $N/n \sum_{k \in S} y_k$. These results are reported in Table 2 for total employment and Table 3 for total wages.

For sample size $n = 5,000$, the expected sample size in the smallest cell (state E and industry group 5) is 9.07. While this might be a little small for poststratification, the probability that this cell has a sample size less than 2, the minimum size necessary for variance estimation, is 0.0011. In our simulations 9 runs had a cell with sample size less than 2. For this sample size, we also report the empirical behavior of poststratified estimator, excluding the 9 problem cases, using the variance estimate (7.6.5) of Särndal, Swensson and Wretman (1992) and its theoretical behavior using the variance approximation given in (7.6.6) of Särndal *et al.* (1992).

Table 1
Business entities by state and industry group

	industry group					
	1	2	3	4	5	sum
A	5,986 (2.11%)	5,548 (1.96%)	7,712 (2.72%)	3,969 (1.40%)	1,299 (0.46%)	24,514 (8.64%)
B	18,782 (6.62%)	31,572 (11.13%)	22,012 (7.76%)	4,982 (1.76%)	4,504 (1.59%)	81,852 (28.85%)
C	13,518 (4.76%)	13,099 (4.62%)	17,837 (6.29%)	5,610 (1.98%)	3,001 (1.06%)	53,065 (18.70%)
D	30,428 (10.72%)	36,017 (12.69%)	32,541 (11.47%)	10,963 (3.86%)	5,399 (1.90%)	115,348 (40.65%)
E	2,225 (0.78%)	2,020 (0.71%)	3,110 (1.10%)	1,076 (0.38%)	515 (0.18%)	8,946 (3.15%)
sum	70,939 (25.00%)	88,256 (31.11%)	83,212 (29.33%)	26,600 (9.38%)	14,718 (5.19%)	283,725

Table 2
Empirical comparison of 4 estimators of total employment

estimator	bias	st. err.	rt. MSE	rt. est. var.
<i>n</i> = 200				
unweighted (theoretical)	0	1,113,220		
unweighted (empirical)	-1,280	1,068,944	1,068,945	1,059,463
cal. margins	-1,394	1,105,201	1,105,201	1,048,873
cal. cell cts.	-218,751	1,008,436	1,031,889	975,140
raking	-462	1,103,172	1,103,172	1,041,490
exp. cal. cell cts.	-227,578	1,000,154	1,025,719	962,153
<i>n</i> = 1,000				
unweighted (theoretical)	0	497,144		
unweighted (empirical)	-5,435	505,941	505,970	501,144
cal. margins	-6,212	506,239	506,277	498,946
cal. cell cts.	-56,118	493,611	496,790	488,222
raking	-4,854	507,938	507,961	499,237
exp. cal. cell cts.	-58,891	492,939	496,445	487,281
<i>n</i> = 5,000				
unweighted (theoretical)	0	220,751		
unweighted (empirical)	1,516	224,088	224,093	222,034
poststr. (theoretical)	0	220,315		
poststr. (empirical, 9 cases excluded)	1,234	223,225	223,228	221,094
cal. margins	1,649	223,091	223,098	220,833
cal. cell cts.	-8,606	222,170	222,337	220,347
raking	3,632	236,355	236,383	220,606
exp. cal. cell cts.	-10,643	223,472	223,725	220,207

Table 3
Empirical comparison of 4 estimators of total wages (tens of thousands of dollars)

estimator	bias	st. err.	rt. MSE	rt. est. var.
<i>n</i> = 200				
unweighted (theoretical)	0	1,682,571		
unweighted (empirical)	-11,119	1,551,186	1,551,226	1,543,483
cal. margins	-11,474	1,582,383	1,582,425	1,510,413
cal. cell cts.	-214,323	1,451,931	1,467,664	1,413,411
raking	-11,220	1,579,842	1,579,882	1,501,170
exp. cal. cell cts.	-221,435	1,438,810	1,455,750	1,393,246
<i>n</i> = 1,000				
unweighted (theoretical)	0	751,406		
unweighted (empirical)	-2,911	772,495	772,501	768,878
cal. margins	-4,372	776,955	776,968	768,869
cal. cell cts.	-51,649	756,201	757,963	751,384
raking	-4,684	778,302	778,316	769,428
exp. cal. cell cts.	-54,305	754,963	756,913	749,832
<i>n</i> = 5,000				
unweighted (theoretical)	0	333,654		
unweighted (empirical)	2,678	336,057	336,068	337,239
poststr. (theoretical)	0	333,765		
poststr. (empirical, 9 cases excluded)	1,802	335,271	335,276	336,192
cal. margins	2,510	334,910	334,920	336,064
cal. cell cts.	-7,149	333,560	333,637	335,006
raking	-4,679	339,074	339,106	335,230
exp. cal. cell cts.	-9,251	334,365	334,493	334,755

The response variables, total employment and total wages, are strongly skewed right. There is one firm (in state C and industry group 4) whose total employment is more than double the total employment of the next largest firm and many hundreds times the mean employment of the

remaining firms. We repeat this study using a population with this firm removed. The results are presented in Tables 4 and 5. In practice with this population, the sampling would normally sample this firm with certainty (a *self representing unit*) and samples constructed from the

remaining firms. Thus Tables 4 and 5 are perhaps more indicative of the relative performance of these estimators in actual practice.

The samples used for Tables 4 and 5 are identical to those used for Tables 2 and 3 except that if the outlier was

included in the sample, it was replaced by a new observation from the population. This was done to improve the comparability of the results of Tables 4 and 5 with those of Tables 2 and 3.

Table 4
Empirical comparison of 4 estimators of total employment: population with outlier removed

estimator	bias	st. err.	rt. MSE	rt. est. var.
<i>n</i> = 200				
unweighted (theoretical)	0	950,688		
unweighted (empirical)	5,395	975,617	975,632	965,448
cal. margins	5,777	1,019,583	1,019,599	963,314
cal. cell cts.	-211,568	909,070	933,365	877,343
raking	6,688	1,018,383	1,018,405	956,867
exp. cal. cell cts.	-217,810	902,756	928,660	868,797
<i>n</i> = 1,000				
unweighted (theoretical)	0	424,552		
unweighted (empirical)	-8,393	422,116	422,199	414,019
cal. margins	-9,430	418,153	418,259	408,577
cal. cell cts.	-58,808	408,391	412,603	399,961
raking	-8,135	419,938	420,016	408,611
exp. cal. cell cts.	-61,014	407,780	412,320	399,311
<i>n</i> = 5,000				
unweighted (theoretical)	0	188,517		
unweighted (empirical)	702	191,631	191,632	188,089
poststr. (theoretical)	0	187,691		
poststr. (empirical, 9 cases excluded)	563	190,854	190,855	187,180
cal. margins	820	190,662	190,664	186,664
cal. cell cts.	-9,376	189,884	190,115	186,202
raking	2,933	205,924	205,944	186,618
exp. cal. cell cts.	-9,922	189,813	190,072	186,140

Table 5
Empirical comparison of 4 estimators of total wages: population with outlier removed

estimator	bias	st. err.	rt. MSE	rt. est. var.
<i>n</i> = 200				
unweighted (theoretical)	0	1,330,930		
unweighted (empirical)	711	1,341,900	1,341,901	1,334,556
cal. margins	1,256	1,387,484	1,387,485	1,318,285
cal. cell cts.	-201,575	1,225,852	1,242,314	1,194,071
raking	1,473	1,386,978	1,386,979	1,311,353
exp. cal. cell cts.	-206,956	1,217,881	1,235,340	1,184,166
<i>n</i> = 1,000				
unweighted (theoretical)	0	594,370		
unweighted (empirical)	-8,169	587,775	587,832	582,524
cal. margins	-10,093	583,606	583,693	576,251
cal. cell cts.	-56,429	569,158	571,948	563,022
raking	-10,529	584,532	584,626	576,282
exp. cal. cell cts.	-58,435	568,277	571,273	562,061
<i>n</i> = 5,000				
unweighted (theoretical)	0	263,923		
unweighted (empirical)	1,185	266,779	266,782	264,110
poststr. (theoretical)	0	263,339		
poststr. (empirical, 9 cases excluded)	566	265,973	265,973	263,210
cal. margins	991	265,449	265,451	262,556
cal. cell cts.	-8,565	264,126	264,265	261,483
raking	-6,008	271,535	271,602	262,021
exp. cal. cell cts.	-9,070	264,038	264,194	261,394

Examining Tables 2 and 3, we see that the $P > Q$ methods, that is those that calibrate the cross classified cell counts using calibration weights which depend upon the marginal classifications, are clearly more biased than the other techniques. However the biases of these estimators relative to their standard deviations decrease with increasing sample sizes. We will show in the next section, that under a highly simplified model, the bias has order n^{-1} and the standard deviation has order $n^{-1/2}$. Consider, for example, the results for the “calibrated cell counts” estimator in Table 2. In this case, the bias divided by the standard error is 0.217, 0.114, 0.039 for $n = 200, 1,000, 5,000$ respectively. For these values of n , the values of $n^{-1/2}$ are 0.0707, 0.0316, 0.0141 and it appears that the former series of three ratios is approximately 3 times the latter series.

It also appears that the exponential back link function f performs slightly better than the linear choice for f . Computationally the former is much more expensive than the latter. We also notice that as the sample sizes increase, the estimators’ performances appear to converge. This is to be expected: because there is no nonresponse, as $n \rightarrow \infty$, $\hat{\beta} \rightarrow 0$, so that the adjustment weights $w = f(\mathbf{z}^T \hat{\beta}) \rightarrow 1$.

Comparing the linear calibrated cell counts estimator to the empirical values of the unweighted estimator, the former is approximately 7.3% more efficient in MSE when $n = 200$ for total employment and 11.7% more efficient for total wages. (This means, for example, that the empirical MSE of the unweighted estimator is 1.117 times the empirical MSE for the linear calibrated cell counts estimator when estimating total wages.) For the exponential calibrated cell counts estimator, the improvement in efficiency relative to the empirical MSE of the unweighted estimator is 8.6% for total employment and 13.5% for total wages. Comparison to the theoretical values for the unweighted estimator would be more favorable to the calibrated cell counts estimators, but we will use the empirical results for the unweighted estimator as the various estimators have all used the same Monte Carlo samples. The calibrated cell counts estimator and exponential calibrated cell counts estimator still have an advantage in MSE over the unbiased estimator at sample size $n = 1,000$.

When the single extreme outlier is removed, leaving 283,724 remaining elements of the population, the calibrated cell count estimators have somewhat better performance relative to the unweighted estimator. For $n = 200$, the linear calibrated cell count estimator offers a 9.3% improvement in efficiency for total employment and a 16.7% improvement for total wages. The comparable ratios for the exponential calibrated cell count estimator are 10.4% for total employment and 18.0% for total wages.

Finally, the variance estimator in equation (15) has a slight downward bias.

4. Model based bias and variance of calibrated estimators

Kott and Chang (2010) derived the asymptotic properties of $\hat{t}_{y,zfV}$ under a different, model-based, probability structure. In that paper \mathcal{S} is a sample selected with selection probabilities d_k^{-1} so that nonresponse is not an issue in \mathcal{S} . Rather, if P the number of benchmark variables \mathbf{x} equals Q the number of model variables \mathbf{z} , Kott and Chang (2010) assume a *prediction* model

$$y_k = \mathbf{x}_k^T \theta + \varepsilon_k, k \in \mathcal{U}. \quad (6)$$

Here θ is a unknown fixed vector, ε_k are model independent errors subject to

$$E(\varepsilon_k | \mathbf{z}_j, I_j, j \in \mathcal{U}) = 0, \quad (7)$$

and I_k is a random variable defined by $I_k = 1$ if $k \in \mathcal{S}$ and $I_k = 0$ otherwise.

When $P > Q$, the model equation (6) must be replaced by

$$y_k = (\mathbf{A}_\infty \mathbf{x}_k)^T \theta + \varepsilon_k, k \in \mathcal{U} \quad (8)$$

for some limiting $Q \times P$ matrix \mathbf{A}_∞ (which is defined in a suitable asymptotic framework, see Kott and Chang (2010)).

Thus when \mathbf{x} represents indicator variables for the complete $I \times J$ cross classification, we have that $\mathbf{x}_k^T \theta$, for k in the $(i, j)^{\text{th}}$ classification, is the mean value of the response variable over the $(i, j)^{\text{th}}$ classification. Hence, by definition, $E(\varepsilon_k | \mathbf{x}_j, j \in \mathcal{U}) = 0$ and, since \mathbf{z} is a function of \mathbf{x} , the model (6) and (7) automatically holds when the sampling (including nonresponse) is noninformative.

However, in our application of calibration, $P = IJ > Q = I + J - 1$ and the model equation (8) has no a priori reason to hold.

Motivated by Kott and Chang (2010) we examine the behavior of calibrated estimates under the following scenario:

1. The benchmark variables \mathbf{x} are indicator variables for some partition of the population into classes C_r . The model (6) automatically holds where the r^{th} component of θ is the population mean of C_r . Let f_r denote the proportion of the population in C_r and $V_r = \text{Var}(\varepsilon_k | k \in C_r)$. We shall also use the notation $\text{Var}(\mathbf{x}_k)$ for V_r when $k \in C_r$.
2. The sample is a simple random sample of size n chosen *with replacement*.
3. The back link function $f(\eta)$ in the estimator $\hat{t}_{y,zfV}$ of equation (2) is $f(\eta) = 1 + \eta$.

Although these assumptions are unrealistic in practice, the main purpose of this section is to heuristically justify a choice, given in the next section, for the matrix \mathbf{V} . At this point, we no longer place any requirements on \mathbf{z} .

We note that in this situation $E(\varepsilon_k | \mathbf{x}_j, I_j, j \in \mathcal{U}) = 0$. Note that (7) will hold if the components of the model variables \mathbf{z} are functions of \mathbf{x} , that is each component of \mathbf{z} is constant on each class. However if $P > Q$, (8) will generally not hold. In any event, in this section we require neither (7) nor (8).

We let

$$\mu_{\mathbf{x}} = \frac{1}{N} \sum_{j \in \mathcal{U}} \mathbf{x}_j$$

$$\mu_{\mathbf{z}} = \frac{1}{N} \sum_{j \in \mathcal{U}} \mathbf{x}_j \mathbf{z}_j^T$$

and the matrix \mathbf{A}_∞ of equation (8) becomes

$$\mathbf{A}_\infty = \mathbf{V}^{-1} \mu_{\mathbf{z}}.$$

Let $\hat{\mu}_{y,zV} = N^{-1} \hat{t}_{y,zV}$ where $\hat{t}_{y,zV}$ is defined as in (2). We have suppressed the f in the notation $\hat{t}_{y,zV}$ because, in this section, $f(\eta) = 1 + \eta$. Letting \bar{y}_s and $\bar{\mathbf{x}}_s$ denote the indicated sample means and using Kott and Chang (2010)

$$\hat{\beta} = \left(\frac{1}{n} \sum_{j \in S} \mathbf{z}_j \mathbf{x}_j^T \mathbf{A}_\infty \right)^{-1} \left(\frac{1}{n} \sum_{j \in S} \mathbf{z}_j y_j \right) + O_p(n^{-1/2})$$

$$\hat{\mu}_{y,zV} = \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{A}_\infty (\mathbf{A}_\infty^T \mathbf{V} \mathbf{A}_\infty)^{-1} \left(\frac{1}{n} \sum_{j \in S} \mathbf{z}_j y_j \right) + O_p(n^{-1})$$

$$= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{A}_\infty (\mathbf{A}_\infty^T \mathbf{V} \mathbf{A}_\infty)^{-1} \mu_{\mathbf{z}}^T \theta + O_p(n^{-1})$$

$$= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{V}^{-1} \mu_{\mathbf{z}} (\mu_{\mathbf{z}}^T \mathbf{V}^{-1} \mu_{\mathbf{z}})^{-1} \mu_{\mathbf{z}}^T \theta + O_p(n^{-1}). \quad (9)$$

If $\hat{\mu}_{y,zV}$ is bounded, as would occur if $f(\eta) = 1 + \eta$ were modified for large η to prevent large calibration weight adjustments, we would have

$$E(\hat{\mu}_{y,zV}) = E(\bar{y}_s) + O(n^{-1}) = \mu_y + O(n^{-1})$$

$$E(\hat{\mu}_{y,zV} | \bar{\mathbf{x}}_s) = \bar{\mathbf{x}}_s^T \theta + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{V}^{-1} \mu_{\mathbf{z}} (\mu_{\mathbf{z}}^T \mathbf{V}^{-1} \mu_{\mathbf{z}})^{-1} \mu_{\mathbf{z}}^T \theta + O(n^{-1})$$

$$\text{Var}[E(\hat{\mu}_{y,zV} | \bar{\mathbf{x}}_s)] = \frac{1}{n} \theta^T (\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{z}}, \mathbf{V}}) \Sigma_{\mathbf{x}} (\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{z}}, \mathbf{V}})^T \theta + o(n^{-1}),$$

where $\Sigma_{\mathbf{x}}$ is the covariance matrix of \mathbf{x} and

$$\mathbf{P}_{\mu_{\mathbf{z}}, \mathbf{V}} = \mu_{\mathbf{z}} (\mu_{\mathbf{z}}^T \mathbf{V}^{-1} \mu_{\mathbf{z}})^{-1} \mu_{\mathbf{z}}^T \mathbf{V}^{-1}. \quad (10)$$

Now

$$\text{Var}(\hat{\mu}_{y,zV} | \bar{\mathbf{x}}_s) = \text{Var}(\bar{y}_s | \bar{\mathbf{x}}_s) + o(n^{-1})$$

$$= \frac{1}{n^2} \sum_{j \in S} V(\mathbf{x}_j) + o(n^{-1})$$

$$E[\text{Var}(\hat{\mu}_{y,zV} | \bar{\mathbf{x}}_s)] = E[\text{Var}(\bar{y}_s | \bar{\mathbf{x}}_s)] + o(n^{-1})$$

$$= \frac{1}{n} \sum_r f_r V_r + o(n^{-1}).$$

It is easily seen that

$$\text{Var}[E(\bar{y}_s | \bar{\mathbf{x}}_s)] = \frac{1}{n} \theta^T \Sigma_{\mathbf{x}} \theta.$$

Since $\text{Var}(\hat{\mu}_{y,zV}) = \text{Var}[E(\hat{\mu}_{y,zV} | \bar{\mathbf{x}}_s)] + E[\text{Var}(\hat{\mu}_{y,zV} | \bar{\mathbf{x}}_s)]$ and similarly for $\text{Var}(\bar{y}_s)$, $\text{Var}(\hat{\mu}_{y,zV}) < \text{Var}(\bar{y}_s)$ to terms $o(n^{-1})$ when

$$\theta^T (\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{z}}, \mathbf{V}}) \Sigma_{\mathbf{x}} (\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{z}}, \mathbf{V}})^T \theta < \theta^T \Sigma_{\mathbf{x}} \theta. \quad (11)$$

The derivation also establishes that the square bias has an asymptotically trivial contribution to the mean square error of $\hat{\mu}_{y,zV}$.

5. A proposed new weight matrix \mathbf{V}^{-1}

In this section we return to our original benchmark \mathbf{x} and model \mathbf{z} variables. When $\mathbf{V} = \mathbf{I}$, the identity matrix, we see from (10) that $\mathbf{P}_{\mu_{\mathbf{z}}, \mathbf{I}} \theta = \mathbf{P}_{\mu_{\mathbf{z}}, \mathbf{I}} \theta$ is the projection of θ onto the span of the columns of $\mu_{\mathbf{z}}$. The left hand side of (11) will be zero if θ is in this column span.

For simplicity, we will write $\mu_{\mathbf{z}}$ as a singular matrix, of rank $I + J - 1$, with one row for each possible double classification cell (i, j) and one column for each row classification i and each column classification j . Thus, the $(i, j)^{\text{th}}$ row of $\mu_{\mathbf{z}}$ has $f_{ij} = N_{ij}/N$ in the columns corresponding to i and j and zero elsewhere. Thus θ will be in the column span of $\mu_{\mathbf{z}}$ if and only if for each i and j

$$\frac{\theta_{ij}}{f_{ij}} = \alpha_i + \beta_j \quad (12)$$

for some α_i and β_j . In other words, the θ_{ij}/f_{ij} satisfy a two way ANOVA model, without interaction, in the column and row classifications.

Recalling that θ_{ij} represents the mean value of the variable of interest y in the $(i, j)^{\text{th}}$ cell, (12) does not appear to be a very promising approximation to the truth. A more likely approximation would be the usual two way ANOVA model

$$\theta_{ij} = \alpha_i + \beta_j. \quad (13)$$

Suppose we change variables $\tilde{\mathbf{x}} = \mathbf{C}\mathbf{x}$ for some diagonal matrix \mathbf{C} . Note that the rows and columns of \mathbf{C} are doubly indexed by (i, j) and we will let c_{ij} denote the diagonal entry in the $(i, j)^{\text{th}}$ row and column. Let $\tilde{\theta} = \tilde{\mathbf{C}}^{-1} \theta$ so that model (6) can be rewritten as

$$y_k = \tilde{\mathbf{x}}_k^T \tilde{\theta} + \varepsilon_k.$$

Now the matrix $\mu_{\tilde{\mathbf{z}}}$ has $c_{ij} f_{ij}$ in the $(i, j)^{\text{th}}$ row and the columns corresponding to i and j . Now $\tilde{\theta}$ will be in the column span of $\mu_{\tilde{\mathbf{z}}}$ if and only if

$$c_{ij}^{-1}\theta_{ij} = \tilde{\theta}_{ij} = c_{ij}f_{ij}(\alpha_i + \beta_j).$$

Thus (13) is equivalent to $c_{ij} = f_{ij}^{-1/2}$. It is easily checked that

$$\tilde{\theta}^T (\mathbf{I} - \mathbf{P}_{\mu_{xz}, \mathbf{I}}) \Sigma_{\tilde{\theta}} (\mathbf{I} - \mathbf{P}_{\mu_{xz}, \mathbf{I}}) \tilde{\theta} = \theta^T (\mathbf{I} - \mathbf{P}_{\mu_{xz}, \mathbf{V}}) \Sigma_{\theta} (\mathbf{I} - \mathbf{P}_{\mu_{xz}, \mathbf{V}}) \theta$$

when $\mathbf{V} = \mathbf{C}^{-2}$. We thus propose using the diagonal matrix \mathbf{V}_o whose diagonal entries are f_{ij} .

With this choice of \mathbf{V}_o , equation (9) suggests the estimator for simple random sampling

$$\hat{t}_{y,z\mathbf{V}_o} = N\bar{y}_s + (T_x - N\bar{x}_s)^T \mathbf{V}_o^{-1} \hat{\mu}_{xz} (\hat{\mu}_{xz}^T \mathbf{V}_o^{-1} \hat{\mu}_{xz})^{-1} \left(\frac{1}{n} \sum_{k \in S} \mathbf{z}_k y_k \right) \quad (14)$$

where $\hat{\mu}_{xz} = n^{-1} \sum_{k \in S} \mathbf{z}_k y_k^T$. In our case both μ_{xz} and μ_x are known from the N_{ij} , but in the spirit of ratio estimation it is preferable to use $\hat{\mu}_{xz}$ in place of μ_{xz} . This heuristic observation has been demonstrated using simulations (not shown) with the QCEW population.

We shall call the estimator $\hat{t}_{y,z\mathbf{V}_o}$ of equation (14) the *weighted calibrated cell counts estimator*.

Simulations with artificial response variables y , also not shown, demonstrate that when the model (13) holds, then weighted calibrated cell counts estimator $\hat{t}_{y,z\mathbf{V}_o}$ performs markedly better than the other estimators considered here. Table 6 gives statistics for the estimator $\hat{t}_{y,z\mathbf{V}_o}$ for the populations and variables studied in Tables 2 - 5.

Comparing to Tables 2 - 5, we see that in all cases $\hat{t}_{y,z\mathbf{V}_o}$ has the highest bias but the lowest MSE of the estimators considered. For $n = 200$ and the full population, $\hat{t}_{y,z\mathbf{V}_o}$ has a 14.8% gain in efficiency (as measured by MSE) relative to the empirical results for the unbiased estimator when estimating total employment and a 21.1% efficiency gain when estimating total wages. For $n = 200$ and the population with a single extreme outlier deleted, the corresponding gains are 14.2% and 21.7% for total employment and total wages respectively.

The Associate Editor suggested that we compare our estimators to a poststratified estimator using collapsed cells to avoid the problem of empty cells in the sample. We explored this question for sample size $n = 200$ where it is most likely that empty cells will occur. We constructed 14 poststrata. Nine of these poststrata are the nine largest cells in the original data. The other 5 poststrata are A1 and A2; A3, A5, and B4; A4, B5, and C4; C5 and D4; and all cells from state E together with D5. After these combinations, the 5 combined poststrata had sizes that ranged between 4.07%

and 5.06% of the population and the 9 retained original cells had sizes in the range of 4.62% to 11.47%.

Table 6
Empirical statistics for $\hat{t}_{y,z\mathbf{V}_o}$ of equation (14)

n	bias	st. err.	rt. MSE	rt. est. var.
<i>Full population - total employment</i>				
200	-244,749	967,066	997,556	923,492
1,000	-64,839	490,758	495,023	483,550
5,000	-10,767	221,702	221,964	219,408
<i>Full population - total wages</i>				
200	-242,528	1,388,489	1,409,511	1,333,793
1,000	-62,091	752,603	755,160	744,315
5,000	-9,821	332,682	332,827	333,782
<i>Population with outlier deleted - total employment</i>				
200	-236,812	881,844	913,088	842,191
1,000	-67,468	405,215	410,793	396,105
5,000	-11,482	189,501	189,848	185,483
<i>Population with outlier deleted - total wages</i>				
200	-228,441	1,194,922	1,216,562	1,151,417
1,000	-66,765	565,008	568,939	557,676
5,000	-11,138	263,699	263,934	260,768

Unfortunately, the author no longer has access to the QCEW data base. Besides the cell counts in Table 1, the author has only the means, standard deviations, and maximum values by cell. The author constructed a pseudo population using the squares of randomly generated gamma variables. The square gamma variables were constructed to have the same cell means and standard deviations as the cell means and standard deviations in the original data. After doing this, the square gamma variables were rounded upwards to integer values. For these pseudo populations, $T_y = 3,149,491$ for employment and 2,305,273, in tens of thousands of dollars, for wages.

A square gamma distribution was used because the gamma distribution is insufficiently right skewed. Even so, in almost all cells the largest value in the original population exceeded the largest value in the pseudo population. Of course without the original data, we cannot distinguish between right skew and a tendency to produce outliers.

10,000 Monte Carlo samples were constructed were taken for each sample size. The results are shown in Table 7. For the poststratified estimator, 5 of the samples of size 200 had an empty poststratum and these runs were excluded from the results in Table 7.

Table 7
Empirical comparison of 4 estimators

estimator	bias	st. err. total employment	rt. MSE	bias	st. err. total wages	rt. MSE
<i>n</i> = 200						
unweighted	644	1,006,956	1,006,956	-9,970	1,481,450	1,481,483
poststratified	-5,387	1,026,266	1,026,280	-2,149	1,548,833	1,548,834
cal. cell cts.	-224,198	942,164	968,472	-203,531	1,377,823	1,392,775
wtd. cal. cell cts.	-248,937	919,419	952,523	-232,558	1,326,234	1,346,469
<i>n</i> = 1,000						
unweighted	-3,317	445,676	445,687	1,544	679,148	679,150
poststratified	-2,967	448,218	448,228	1,672	685,370	685,372
cal. cell cts.	-54,311	436,821	440,185	-44,942	665,799	667,314
wtd. cal. cell cts.	-63,327	432,396	437,008	-54,913	660,726	663,004
<i>n</i> = 5,000						
unweighted	2,466	206,249	206,264	-2,539	304,852	304,863
poststratified	2,108	205,661	205,672	-2,705	304,751	304,763
cal. cell cts.	-8,265	204,693	204,859	-12,096	303,231	304,472
wtd. cal. cell cts.	-10,551	204,080	204,352	-14,697	302,311	302,668

Evidently the poststratification did not help. Even though no poststratum had an expected count below eight, the actual poststrata had quite variable sizes. In addition, the cell populations are quite skewed so that the poststrata sample means are quite variable.

The other conclusions for the pseudo populations reflect the conclusions from the actual populations. In particular, when *n* = 200 and for the employment pseudo population, the weighted calibrated cell counts estimator \hat{t}_{y,zV_o} has an 11.8% gain in efficiency relative to the unbiased estimator. For the wages pseudo population and *n* = 200, the efficiency gain is 21.1%.

6. Concluding remarks

The use in (3) of weight matrices $V(\beta)^{-1}$ which depend upon β has not been explored in this paper. Experimentation with the use of such a matrix was not encouraging. Computation time increased dramatically, and there were significant numbers of cases which failed to numerically converge, with no improvement in efficiency over the fixed **V** estimators considered here. Perhaps the authors did not try the right $V(\beta)$.

Besides the exponential back link function, the authors tried the logistic back link $f(\eta) = (1 + e^{-\eta})^{-1}$. These runs also did not converge. On reflection, the reason is obvious: because in the simulations there was no nonresponse or noncoverage problems, the calibration weight adjustments $f(z^T\beta) \rightarrow 1$ as *n* \rightarrow ∞ . But 1 is not in the range of *f*. It should be noted that in Chang and Kott (2008) a logistic back link was used to correct for nonresponse.

Several obvious issues arise. For example, how would the results of this study change if a more complicated

sampling design than simple random sampling were used, or if non response and/or non coverage occur and the calibration was used to correct for it. Falk (2010) considers these questions both theoretically and with further simulations using the QCEW population. Falk (2010) also considers non linear link functions.

There are obvious extensions to 3-way (and beyond) cross classified data. If *I, J, K* denote the number of cells in each of the 3 classifications, there are *IJK* fully classified cells whose totals can be used for benchmark **x** variables. There are *IJ* + *IK* + *JK* − *I* − *J* − *K* + 1 one-way and two-way marginal variables that can be used for model **z** variables. Clearly, one might not want to use the plethora of variables available.

In the context of linear calibration using the same **x** and **z** variables, several studies have been made on the choice of variables. Examples of such studies are Banker, Rathwell and Majkowski (1992), Silva and Skinner (1997), and Clark and Chambers (2008). The last paper remarks that too many variables can deteriorate the MSE of \hat{T}_y .

The alternatives to poststratification discussed here can be used in the presence of small and even empty cells. For example, in our simulations, the expected count in the state E, industry group 5, cell is 0.36 when *n* = 200. One might be tempted to collapse cells and use poststratification. Generally, however, it is not possible to do so and maintain the convenient doubly classified structure of the data. Our approaches, like poststratification, introduce weights for the purpose of sample balancing but avoid collapsing cells. These approaches generally increase bias but can offer substantial reductions in MSE.

Furthermore, in the presence of nonresponse or non-coverage, the inverse of the weight adjustments can be

considered, under a quasi-randomization model for the response or coverage, as estimated probabilities of response and/or coverage. In our calibration approaches, these probabilities are assumed to be a function of the row and column classifications. When cells are collapsed without maintaining the double classification, these probabilities are harder to interpret.

Acknowledgements

The author would like to thank Phil Kott and John Eltinge for several very interesting insights. The author also thanks Larry Huff and Ken Robertson of the Bureau of Labor Statistics for help in obtaining and understanding the data.

Appendix

Here we derive, using Chang and Kott (2008) equations (16) and (17), sample based variance estimators for the 4 estimators studied in Section 2.

Let

$$\hat{\mathbf{H}}_y = \frac{\partial \hat{f}_{y,zfV}}{\partial \beta}(\hat{\beta}).$$

Here $\hat{f}_{y,zfV}$ is defined in (2). $\hat{\mathbf{H}}_y$ is a row vector with one entry for each \mathbf{z} variable. In out case, $\hat{\mathbf{H}}_y$ has $(I + J - 1)$ entries, one for each of the $I + J - 1$ linearly independent indicator variables for the row and column classifications.

For the calibrated margins and calibrated cell counts estimators, $f(\eta) = 1 + \eta$. Define the constants s_{ij} and t_{ij} by

$$s_{ij} = \frac{N}{n} r_{ij}$$

$$t_{ij} = \frac{N}{n} \sum_{k \in S_{ij}} y_k.$$

Then a simple calculation shows that if a entry exists in $\hat{\mathbf{H}}_y$ for the i^{th} row classification, we place in that entry $\sum_j t_{ij}$. Similarly if a entry exists for the j^{th} column classification, we place in that entry $\sum_i t_{ij}$. Here we use the convention that if the i^{th} row or j^{th} column is not one of the chosen $I + J - 1$ linearly independent indicator variables then corresponding β_i or β_j is 0.

For the raking and exponential calibrated cell counts estimators, $f(\eta) = e^\eta$ and we can similarly calculate $\hat{\mathbf{H}}_y$ using instead

$$s_{ij} = \frac{N}{n} \exp(\hat{\beta}_i + \hat{\beta}_j) r_{ij}$$

$$t_{ij} = \frac{N}{n} \sum_{k \in S_{ij}} \exp(\hat{\beta}_i + \hat{\beta}_j) y_k.$$

Here we use the convention that if the i^{th} row or j^{th} column is not one of the chosen $I + J - 1$ linearly independent indicator variables then corresponding β_i or β_j is 1.

Analogously to (2), let

$$\hat{f}_{\mathbf{x}zfV} = \sum_{k \in S} d_k f(\mathbf{z}_k^T \hat{\beta}) x_k.$$

$\hat{f}_{\mathbf{x}zfV}$ is a column vector with one entry for each \mathbf{x} variable. Define the $\hat{\mathbf{H}}$ matrix to be

$$\hat{\mathbf{H}} = \frac{\partial \hat{f}_{\mathbf{x}zfV}}{\partial \beta}(\hat{\beta}).$$

$\hat{\mathbf{H}}$ is a matrix with one row for each \mathbf{x} variable and one column for each \mathbf{z} variable.

For the calibrated cell counts and exponential calibrated cell counts estimators the matrix $\hat{\mathbf{H}}$ has dimensions $IJ \times (I + J - 1)$. Each of the rows of $\hat{\mathbf{H}}$ corresponds to a pair (i, j) of row and column classifications. We place s_{ij} in the row corresponding to (i, j) and the columns corresponding to the i^{th} row classification and the j^{th} column classification (whenever these columns exist). All other entries of $\hat{\mathbf{H}}$ are set to zero.

For the calibrated margins and raking estimators the matrix $\hat{\mathbf{H}}$ has dimensions $(I + J - 1) \times (I + J - 1)$. If a row (and hence a column) of $\hat{\mathbf{H}}$ exists for the i^{th} row classification we put $\sum_j s_{ij}$ in the corresponding diagonal entry of $\hat{\mathbf{H}}$. Similarly, if a row and column exist for the j^{th} column classification, we put $\sum_i s_{ij}$ on the diagonal of $\hat{\mathbf{H}}$. We place s_{ij} in the entry whose row corresponds to the i^{th} row classification and whose column corresponds to j^{th} column classification (whenever these exist). We also place s_{ij} in the entry whose column corresponds to the i^{th} row classification and whose row corresponds to j^{th} column classification (again whenever these exist). All other entries of $\hat{\mathbf{H}}$ are set to zero.

Let $\mathbf{B} = \hat{\mathbf{H}}_y^T (\hat{\mathbf{H}}^T \mathbf{V}^{-1} \hat{\mathbf{H}})^{-1} \hat{\mathbf{H}}^T \mathbf{V}^{-1}$ where currently we are using an identity matrix for \mathbf{V} . \mathbf{B} has dimensions $1 \times (I + J - 1)$ for the calibrated margins and raking estimators and $1 \times IJ$ for the calibrated cell counts and the exponential calibrated cell counts estimators. In the former cases, we will denote the entries of \mathbf{B} by b_i or b_j , and, for the single case when a column or row index does not correspond to one of the $I + J - 1$ independent indicator variables, we will set the corresponding b to zero. In the latter cases, we will denote the entries of \mathbf{B} by b_{ij} . For $k \in S_{ij}$, let $u_k = w_{ij}(y_k - b_i - b_j)$ for the calibrated margins and raking estimators and $u_k = w_{ij}(y_k - b_{ij})$ for the calibrated cell counts and exponential calibrated cell counts estimators.

Essentially Chang and Kott (2008) showed that, asymptotically, the calibrated estimator has the same form as a regression estimator of the form Särndal *et al.* (1992)

equation (6.6.1) where the above \mathbf{B} plays the role of \mathbf{B} in (6.6.1) and the sampling weights d_k are replaced by $d_k f(\mathbf{z}_k^T \hat{\beta})$. For non replacement designs, they propose to estimate the variance of $\hat{t}_{y, \mathbf{z} f \mathbf{V}}$ using the analogous changes to Särndal *et al.* (1992) equation (6.6.3).

For simple random sampling, and in the absence of nonresponse or noncoverage, the variance estimator works out to

$$\hat{\mathbf{V}} = \frac{N^2}{n} (1 - n/N) s_u^2 \quad (15)$$

where s_u^2 is the sample variance of the u_k .

In the presence of nonresponse, if one assumes that the respondents \mathcal{S} are a Poisson sample from the original simple random sample with Poisson probabilities $f(\mathbf{z}^T \beta_0)^{-1}$, the variance estimator becomes

$$\hat{\mathbf{V}} = \frac{N^2}{n} (1 - n/N) s_u^2 + \frac{N}{n} \sum_i \sum_j (1 - w_{ij}) \sum_{k \in S_{ij}} u_k^2 \quad (16)$$

where s_u^2 is the sample variance of the u_k . The same formula works for noncoverage where $f(\mathbf{z}^T \beta_0)^{-1}$ represents the combined coverage and response probability in a three stage model in which the covered universe is assumed to be a Poisson sample from the desired universe, the sample is a simple random sample from the covered universe, and the respondents are a Poisson sample from the original sample.

References

- Banker, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 canadian census. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 764-769.
- Chang, T., and Kott, S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 557-571.
- Clark, R., and Chambers, R.L. (2008). Adaptive Calibration for Prediction of Finite Population Totals. University of Wollongong (on line working paper).
- Falk, G. (2010). *Calibration Adjustment for Nonresponse in Cross-Classified Data*, University of Virginia (dissertation).
- Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the Presence of Nonresponse with application to the 1987-1988 Nationalwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.
- Kott, S., and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105, 1265-1275.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Silva, P.L.D.N., and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 23-32.

On variances of changes estimated from rotating panels and dynamic strata

Paul Knottnerus and Arnout van Delden¹

Abstract

Many business surveys provide estimates for the monthly turnover for the major Standard Industrial Classification codes. This includes estimates for the change in the level of the monthly turnover compared to 12 months ago. Because business surveys often use overlapping samples, the turnover estimates in consecutive months are correlated. This makes the variance calculations for a change less straightforward. This article describes a general variance estimation procedure. The procedure allows for yearly stratum corrections when establishments move into other strata according to their actual sizes. The procedure also takes into account sample refreshments, births and deaths. The paper concludes with an example of the variance for the estimated yearly growth rate of the monthly turnover of Dutch Supermarkets.

Key Words: Births; Business surveys; Conditional covariances; Deaths; Overlapping samples; Stratum corrections.

1. Introduction

In many surveys a changing population is repeatedly sampled so that the level and the change in the level of a characteristic between two occasions can be estimated. For example, in many countries a monthly business survey is held to estimate the level of the monthly turnover and the change in that level compared to a month or a year ago; see Konschnik, Monsour and Detlefsen (1985). Another example is the labour force survey in which the population is sampled on a monthly basis to estimate the number of unemployed persons and the unemployment rate. Variance estimation is needed to judge whether the observed changes are statistically significant. Variance estimation is also needed in the design stage of the survey, to determine the optimal sample size and allocation or to determine the optimal estimator.

In repeated surveys, changes are often estimated by using a stratification of the population. Businesses are extremely heterogeneous in terms of size and type of economic activity. Therefore, business surveys are usually designed as a stratified simple random sample selected without replacement (STSRs); see Smith, Pont and Jones (2003). In surveys for households or individuals the sample is usually not stratified because households are less heterogeneous. Some social surveys, such as labour force surveys, however, use poststratification to reduce the variance and bias of the estimator.

In deriving formulas for the variance of an estimated change in a population with dynamic strata, one has to pay attention to three complicating factors. Firstly, the change in a level is the result of two components. One component is due to the change in the population mean of units that remain in the same stratum on both occasions. The other

component is caused by the change in the stratum composition between two occasions resulting from births and deaths in the population and from population units that migrate between strata; see Holt and Skinner (1989). Secondly, due to the migration of population units between strata, the estimated mean of stratum h at occasion t may be correlated with the mean of stratum ℓ at occasion $t + 1$. Thirdly, another complicating factor is that the population is repeatedly sampled, resulting in partially overlapping samples between two occasions. Different rotating panel designs may be used in business surveys.

Various authors have derived formulas for design-based variance estimators for the estimation of changes. Assuming a large population without births and deaths, Kish (1965) derived an expression for the variance of an estimated change based on overlapping samples. Tam (1984) removed the assumption of a large population. Elaborating on Tam's results, Qualité and Tillé (2008) compare several variance estimators of an estimated change. Wood (2008) generalizes Tam's results for surveys with unequal probabilities. Lowerre (1979) and Laniel (1987) deal with the variance estimation of a change in dynamic populations, but they do not take stratification into account. Hidioglou, Särndal and Binder (1995) deal with dynamic populations and stratification, but not with changing strata. Nordberg (2000) and Berger (2004) derived formulas for the more complicated situation of a dynamic population with units that move between strata. For the Swedish sampling design Nordberg (2000) derives formulas using inclusion indicators which requires some algebra. Assuming that the size of the overlap of two samples at two different occasions is fixed, Berger (2004) derives his formulas based on Poisson sampling conditional on the sample size per stratum which requires some matrix algebra.

1. Paul Knottnerus and Arnout van Delden, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: pkts@cbs.nl and adln@cbs.nl.

In this paper, we derive the expressions for STSRS sampling in a more straightforward manner without assuming that sizes of overlaps are fixed. Furthermore, unlike the Swedish design, the Dutch one doesn't require time-consuming calculations for estimating one of the variance components for a change. In addition, we propose an alternative estimation method for sampling designs with such a non-zero component. In order to clarify the variance estimation procedure, we describe its application to the yearly growth rates of the turnover of Dutch Supermarkets of 4-week periods.

The outline of the paper is as follows. Section 2 briefly describes the Dutch business survey for monthly turnover, including the sampling design. The variance formulas for the estimator of a change are derived in section 3. Section 4 illustrates the variance estimation procedure by comparing the variances of two different estimators for the yearly growth rate of the monthly turnover of Dutch Supermarkets in the period 2003-2004. Section 5 summarizes the main results and conclusions.

2. The sampling design of the Dutch business surveys

Every month Statistics Netherlands estimates the monthly turnover for some of the major SIC codes. The publication includes the 12-month growth rates of the monthly turnover, *i.e.*, the relative change in the monthly level of turnover compared to 12 months ago. Throughout this paper we will refer to this growth rate as the yearly growth rate.

All statistical units or *establishments* are listed in the General Business Register (GBR) that is maintained by Statistics Netherlands. The register is updated each month for births and deaths from administrative sources, while once a year, on December 31, the size category and the type of economic activity (SIC code) are updated. Note that the registration in the GBR may lag behind the changes in the population (births, deaths, size class changes *etc.*). Moreover, the (unknown) deaths in the frame may lead to a biased estimate of the level of the turnover. In order to avoid this kind of bias, it is important to quickly detect and remove deaths from the frame. Deaths detected in the sample may play a role here. However, a further analysis and correction of these errors are beyond the scope of this paper on variance estimation for growth rates. For estimating these variances, we assume that the population units and their characteristics in the register are correct. Likewise, we assume that there is zero non-response among the surveys.

Every first day of the month an STSRS-like sample from the GBR is conducted to estimate the turnover of the current month. In fact, a rotating sample is used. The sample is stratified by size and by type of economic activity. The

actual probability of selection depends on size and economic activity. The probability of selection increases with the size of establishment, with the largest establishments being included in the sample with probability 1. For some SICs there are not only survey data available but also data from administrative sources. The units already present in the administrative files are considered as a separate stratum. The estimates from this stratum have a zero variance.

The sample is updated in two ways. Every month the sample is updated to correct for births and deaths in the population. Once a year, in January, 10% of the sample units are replaced and stratum corrections are carried out. We will discuss the monthly and yearly updates in more detail.

2.1 Monthly update

Each month t ($t = 1, 2, \dots$) a fixed proportion f_h of the N_h^t units in stratum U_h^t is sampled ($h = 1, \dots, H$). This results in a sample s_h^t of size $n_h^t = f_h N_h^t$. Hence, the actual number of units in the sample may change from month to month due to births and deaths in the population. Note that apart from minor round-off errors the actual sampling fraction f_h does not depend on month t . In fact, the update procedure for s_h^t in month t is as follows. Define $U_{0h}^{t-1,t}$ as the set of births in stratum h in month $t-1$ and denote its size by $N_{0h}^{t-1,t}$. The number of sampled units from $U_{0h}^{t-1,t}$ in month t is $n_{0h}^{t-1,t} = f_h N_{0h}^{t-1,t}$. In addition, denote the further required difference $n_h^t - n_{0h}^{t-1,t}$ by $n_{h,REQ}^{t-1,t}$ and define $s_{h,PRE}^t$ by $s_{h,PRE}^t = s_h^{t-1} \cap U_h^t$, that is the set of units in s_h^{t-1} that still exist in month t . Let $n_{h,PRE}^t$ denote the size of $s_{h,PRE}^t$. When $n_{h,PRE}^t \geq n_{h,REQ}^{t-1,t}$, randomly drop the difference, otherwise select the difference from $U_h^t \setminus U_{0h}^{t-1,t} \setminus s_{h,PRE}^t$. Note that units dropped from the sample in month $t-1$ or earlier may be re-selected in month t .

2.2 Yearly update

Each January, the sample is updated to account for both a re-stratification of the units and a sample replacement of 10%. All sample units of December that still exist in January are stratified according to their actual size, *i.e.*, the number of employees and the SIC-code of January. The size class boundaries themselves remain unchanged. Consequently, the resulting sample from a stratum according to the new January stratification may consist of units with different inclusion probabilities because units move between strata with different sampling fractions.

In order to correct for possibly different inclusion probabilities in stratum ℓ , denote the substratum consisting of units that belonged to stratum h in December and in January to stratum ℓ by $U_{h\ell}^{dec,jan}$ and denote its size by $N_{h\ell}^{dec,jan}$ ($h, \ell = 1, \dots, H$). In analogy with the monthly update procedure define $s_{h\ell,PRE}^{jan}$ by $s_{h\ell,PRE}^{jan} = s_h^{dec} \cap U_{h\ell}^{dec,jan}$ and let $n_{h\ell,PRE}^{jan}$ denote the size of $s_{h\ell,PRE}^{jan}$. Since the required size of

sample $s_{h\ell, \text{REQ}}^{\text{dec, jan}}$ from $U_{h\ell}^{\text{dec, jan}}$ in January is $n_{h\ell, \text{REQ}}^{\text{dec, jan}} = f_t N_{h\ell}^{\text{dec, jan}}$, the yearly update of sample $s_{h\ell, \text{PRE}}^{\text{jan}}$ is carried out as follows.

Firstly, when $n_{h\ell, \text{PRE}}^{\text{jan}} \geq n_{h\ell, \text{REQ}}^{\text{dec, jan}}$, randomly drop the difference from $s_{h\ell, \text{PRE}}^{\text{jan}}$. In addition, 10% of the $n_{h\ell, \text{REQ}}^{\text{dec, jan}}$ remaining units in $s_{h\ell, \text{PRE}}^{\text{jan}}$ is replaced by units from $U_{h\ell}^{\text{dec, jan}} \setminus s_{h\ell, \text{PRE}}^{\text{jan}}$ provided that the latter set contains enough units. When there are not enough units available, the number of replaced units is only $N_{h\ell}^{\text{dec, jan}} - n_{h\ell, \text{PRE}}^{\text{jan}}$. Secondly, when $n_{h\ell, \text{PRE}}^{\text{jan}} < n_{h\ell, \text{REQ}}^{\text{dec, jan}}$, select the difference from $U_{h\ell}^{\text{dec, jan}} \setminus s_{h\ell, \text{PRE}}^{\text{jan}}$. Subsequently, an additional replacement of $n_{h\ell, \text{PRE}}^{\text{jan}} - 0.9n_{h\ell, \text{REQ}}^{\text{dec, jan}}$ units in $s_{h\ell, \text{PRE}}^{\text{jan}}$ takes place when this difference is positive and enough new units are available. This procedure is done for all substrata $h\ell$, including $h = \ell$. Thirdly, similar to the monthly update procedure the number of sampled units in January from substratum $U_{0\ell}^{\text{dec, jan}}$ of new births in stratum ℓ is $n_{0\ell}^{\text{dec, jan}} = f_t N_{0\ell}^{\text{dec, jan}}$. In addition, note that this approach can also be followed when class size boundaries or sampling fractions are changed in January.

Apart from the stratum corrections in January, the resulting sample in month t can be considered more or less as a set of SRS samples from the strata U_h^t . When the population and the strata h are stable over the years, the procedure described so far amounts to a standard STSRS sampling design for month t . Therefore, Statistics Netherlands uses the familiar variance formulas for the STSRS sampling design for estimating the variance of the level of the monthly turnover. In the next section we show how the variance for a change of the level can be estimated under such an STSRS assumption.

3. Variance of the yearly growth rate of monthly turnover

3.1 Variance of the yearly growth rate

Let O^t denote the total turnover of all establishments in the population in month t and $g^{t,s}$ the relative change in the level of turnover between months t and s , i.e.,

$$g^{t,s} = \frac{O^t}{O^s} - 1 \quad (t > s).$$

For the corresponding estimates it holds by definition that

$$\hat{g}^{t,s} = \frac{\hat{O}^t}{\hat{O}^s} - 1, \quad (1)$$

where a “hat” indicates an estimate; for an estimator we use the same notation. Furthermore, define

$$G^{t,t-12} \equiv \frac{O^t}{O^{t-12}} = 1 + g^{t,t-12}.$$

In order to estimate the variance of the yearly growth rate of the monthly turnover, we use the first-order Taylor series expansion of a ratio of two estimators. That is,

$$\begin{aligned} \text{var}(\hat{g}^{t,t-12}) &= \text{var}\left\{\frac{\hat{O}^t}{\hat{O}^{t-12}}\right\} \\ &\approx \frac{\text{var}(\hat{O}^t - G^{t,t-12}\hat{O}^{t-12})}{(O^{t-12})^2} \\ &= \frac{\text{var}(\hat{O}^t) + (G^{t,t-12})^2 \text{var}(\hat{O}^{t-12}) - 2G^{t,t-12} \text{cov}(\hat{O}^{t-12}, \hat{O}^t)}{(O^{t-12})^2}. \quad (2) \end{aligned}$$

The major problem is the estimation of $\text{cov}(\hat{O}^{t-12}, \hat{O}^t)$. In the next sections we examine this term and its estimation.

3.2 The covariance term of the yearly growth rate

Using the stratified sampling design, we can write $\text{cov}(\hat{O}^{t-12}, \hat{O}^t)$ from (2) as

$$\begin{aligned} \text{cov}(\hat{O}^{t-12}, \hat{O}^t) &= \text{cov}\left(\sum_{h=1}^H N_h^{t-12} \bar{o}_h^{t-12}, \sum_{\ell=1}^H N_\ell^t \bar{o}_\ell^t\right) \\ &= \sum_{h=1}^H \sum_{\ell=1}^H N_h^{t-12} N_\ell^t \text{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^t), \quad (3) \end{aligned}$$

where \bar{o}_h^{t-m} stands for the sample mean of the turnover in stratum h in month $t-m$ ($m = 0, 12$). Note that the stratification of the units in month $t-12$ may differ from that in month t . As we have seen in section 2.2, the standard refreshment of the panel takes place in January. Furthermore, each establishment is allocated to the correct stratum h according to its actual number of employees in January ($h = 1, \dots, H$). To take these design features into account, define

- $N_{h\ell}^{t-12,t}$: size of substratum $U_{h\ell}^{t-12,t}$, i.e., the set of units that in month $t-12$ belonged to stratum h and in month t to stratum ℓ ($h, \ell = 1, \dots, H$);
- $O_{h\ell}^{t-m}$: the substratum population total of the turnover in $U_{h\ell}^{t-12,t}$ in month $t-m$ ($m = 0, 12$);
- $\bar{O}_{h\ell}^{t-m}$: the substratum population mean of the turnover in $U_{h\ell}^{t-12,t}$ in month $t-m$ [i.e., $\bar{O}_{h\ell}^{t-m} = O_{h\ell}^{t-m} / N_{h\ell}^{t-12,t}$ ($m = 0, 12$)];
- $n_{h\ell}^{t-m}$: size of sample $s_{h\ell}^{t-m}$, i.e., the actual sample from $U_{h\ell}^{t-12,t}$ in month $t-m$ ($0 \leq m \leq 12$);
- $o_{h\ell}^{t-m}$: the sample total of the turnover in $s_{h\ell}^{t-m}$ ($m = 0, 12$);
- $\bar{o}_{h\ell}^{t-m}$: the sample mean of the turnover in $s_{h\ell}^{t-m}$ [i.e., $\bar{o}_{h\ell}^{t-m} = o_{h\ell}^{t-m} / n_{h\ell}^{t-m}$ ($m = 0, 12$)];
- $n_{h\ell}^{t-12,t}$: number of units in the overlap $s_{h\ell}^{t-12,t} \equiv s_{h\ell}^{t-12} \cap s_{h\ell}^t$.

$\bar{o}_{h\ell, \text{OLP}}^{t-m}$: the sample mean of the turnover in the overlap
 $s_{h\ell}^{t-12,t}$ in month $t-m$. [i.e., $\bar{o}_{h\ell, \text{OLP}}^{t-m} = o_{h\ell, \text{OLP}}^{t-m} / n_{h\ell}^{t-12,t}$ ($m = 0, 12$)].

In addition to the notation in section 2, define the auxiliary stratum 0 for the *births* in months $t-12, \dots, t-1$ and likewise, stratum $H+1$ for the *deaths* in that period. Then \bar{o}_h^{t-12} and \bar{o}_ℓ^t can be written as

$$\bar{o}_h^{t-12} = \sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{t-12}$$

$$\bar{o}_\ell^t = \sum_{k=0}^H \frac{n_{k\ell}^t}{n_\ell^t} \bar{o}_{k\ell}^t,$$

respectively ($1 \leq h, \ell \leq H$). Consequently, the covariances in (3) can be rewritten as

$$\text{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^t) = \text{cov}\left(\sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{t-12}, \sum_{k=0}^H \frac{n_{k\ell}^t}{n_\ell^t} \bar{o}_{k\ell}^t\right) \quad (4a)$$

$$= \frac{1}{n_h^{t-12} n_\ell^t} \text{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t), \quad (4b)$$

where we used $\text{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{k\ell}^t \bar{o}_{k\ell}^t) = 0$ ($k \neq h$) and $\text{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t) = 0$ ($g \neq \ell$). The latter covariance is zero because

$$\begin{aligned} \text{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t) &= E \text{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t | n_{hg}^{t-12}, n_{h\ell}^t) \\ &\quad + \text{cov}\{E(n_{hg}^{t-12} \bar{o}_{hg}^{t-12} | n_{hg}^{t-12}, n_{h\ell}^t), E(n_{h\ell}^t \bar{o}_{h\ell}^t | n_{hg}^{t-12}, n_{h\ell}^t)\} \\ &= 0 + \bar{o}_{hg}^{t-12} \bar{o}_{h\ell}^t \text{cov}(n_{hg}^{t-12}, n_{h\ell}^t) = 0. \end{aligned}$$

In the last line we also used that for $1 \leq g \leq H+1$

$$\text{cov}(n_{hg}^{t-12}, n_{h\ell}^t) = 0. \quad (5)$$

For a justification and the underlying assumptions of (5), see Appendix A. Moreover, in Appendix A we propose an alternative estimation method when this covariance is non-negligible. The covariance in (4b) can be expressed as

$$\begin{aligned} \text{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t) &= E\{\text{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t | v_{h\ell})\} \\ &\quad + \text{cov}\{E(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12} | v_{h\ell}), E(n_{h\ell}^t \bar{o}_{h\ell}^t | v_{h\ell})\} \quad (6) \end{aligned}$$

where $v_{h\ell} = (n_{h\ell}^{t-12}, n_{h\ell}^{t-12,t}, n_{h\ell}^t)$. The first component on the right-hand side is

$$\begin{aligned} E\{\text{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t | v_{h\ell})\} &= E\{n_{h\ell}^{t-12} n_{h\ell}^t \text{cov}(\bar{o}_{h\ell}^{t-12}, \bar{o}_{h\ell}^t | v_{h\ell})\} \\ &= E\left\{n_{h\ell}^{t-12} n_{h\ell}^t \left(\frac{n_{h\ell}^{t-12,t} / n_{h\ell}^{t-12}}{n_{h\ell}^t} - \frac{1}{N_{h\ell}^{t-12,t}}\right) S_{h\ell}^{t-12,t}\right\}. \quad (7) \end{aligned}$$

In the last line we used (26) in Appendix B. Furthermore,

$$S_{h\ell}^{t-12,t} = \frac{1}{N_{h\ell}^{t-12,t} - 1} \sum_{i=1}^{N_{h\ell}^{t-12,t}} (o_{h\ell i}^{t-12} - \bar{o}_{h\ell}^{t-12})(o_{h\ell i}^t - \bar{o}_{h\ell}^t). \quad (8)$$

The second component on the right-hand side of (6) is equal to $\bar{o}_{h\ell}^{t-12} \bar{o}_{h\ell}^t \text{cov}(n_{h\ell}^{t-12}, n_{h\ell}^t) = 0$ on account of (5). It therefore follows from (4) and (6) that

$$\begin{aligned} \text{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^t) &= \\ E\left\{n_{h\ell}^{t-12} n_{h\ell}^t \left(\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^t} - \frac{1}{N_{h\ell}^{t-12,t}}\right) S_{h\ell}^{t-12,t}\right\}. \quad (9) \end{aligned}$$

3.3 Estimation of the covariance term of the yearly growth rate

Expression (9) can be estimated from the overlapping sample $S_{h\ell}^{t-12,t}$ by

$$\hat{\text{cov}}(\bar{o}_h^{t-12}, \bar{o}_\ell^t) = \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_h^{t-12} n_\ell^t} \left(\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^t} - \frac{1}{N_{h\ell}^{t-12,t}}\right) \hat{S}_{h\ell, \text{OLP}}^{t-12,t}, \quad (10)$$

where

$$\hat{S}_{h\ell, \text{OLP}}^{t-12,t} = \frac{1}{n_{h\ell}^{t-12,t} - 1} \sum_{i=1}^{n_{h\ell}^{t-12,t}} (o_{h\ell i}^{t-12} - \bar{o}_{h\ell, \text{OLP}}^{t-12})(o_{h\ell i}^t - \bar{o}_{h\ell, \text{OLP}}^t).$$

Note that (10) is unbiased for estimating (9) because

$$E(\hat{S}_{h\ell, \text{OLP}}^{t-12,t} | v_{h\ell}) = S_{h\ell}^{t-12,t}.$$

Although (10) results in reasonable estimates for sufficiently large $n_{h\ell}^{t-12,t}$, a disadvantage of the covariance estimator $\hat{S}_{h\ell, \text{OLP}}^{t-12,t}$ in (10) is that for small $n_{h\ell}^{t-12,t}$ it may lead to a negative estimate of $\text{var}(\hat{O}^t - G^{t,t-12} \hat{O}^{t-12})$ in the numerator of (2). Recall that this variance is estimated by

$$\begin{aligned} \hat{\text{var}}(\hat{O}^t - G^{t,t-12} \hat{O}^{t-12}) &= \hat{\text{var}}(\hat{O}^t) + (\hat{G}^{t,t-12})^2 \hat{\text{var}}(\hat{O}^{t-12}) \\ &\quad - 2\hat{G}^{t,t-12} \hat{\text{cov}}(\hat{O}^t, \hat{O}^{t-12}). \quad (11) \end{aligned}$$

Therefore, we propose an alternative estimator to $\hat{S}_{h\ell, \text{OLP}}^{t-12,t}$ in (10). Define the standard deviations

$$\hat{S}_{h\ell}^{t-m} = \sqrt{\frac{1}{n_{h\ell}^{t-m} - 1} \sum_{i=1}^{n_{h\ell}^{t-m}} (o_{h\ell i}^{t-m} - \bar{o}_{h\ell}^{t-m})^2} \quad (m = 0, 12).$$

We propose the following modified estimator for $S_{h\ell}^{t-12,t}$

$$\hat{S}_{h\ell}^{t-12,t} = \hat{\rho}_{h\ell, \text{OLP}}^{t-12,t} \hat{S}_{h\ell}^{t-12} \hat{S}_{h\ell}^t, \quad (12)$$

where $\rho^{t-12,t}$ is the correlation between the variables o^t and o^{t-12} in $U_{h\ell}^{t-12,t}$ and $\hat{\rho}_{h\ell, \text{OLP}}^{t-12,t}$ is its estimate from $s_{h\ell}^{t-12,t}$. According to (10) and (12) covariance (3) can be estimated by

$$\begin{aligned} \hat{\text{cov}}(\hat{O}^{t-12}, \hat{O}^t) &= \\ \sum_{h=1}^H \sum_{\ell=1}^H \frac{N_{h\ell}^{t-12} N_{h\ell}^t}{n_{h\ell}^{t-12} n_{h\ell}^t} n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right) \hat{S}_{h\ell}^{t-12,t}. \quad (13) \end{aligned}$$

For the estimate $\hat{\rho}_{h\ell, \text{OLP}}^{t-12,t}, \hat{\rho}_{h\ell, \text{OLP}}^{t-12,t} \leq 1$ always holds whereas using (10) may lead implicitly to an estimated correlation larger than 1 and a possibly negative outcome of (11). See the next section for an example. In all applications met so far, negative outcomes of (11) could be explained by the fact that unlike (12) use of (10) leads implicitly to an estimated correlation larger than 1. This is in line with the findings of Berger (2004, page 462) that an overestimation of the correlation between \hat{O}^{t-12} and \hat{O}^t may lead to a serious underestimation of the variance of a change. Nevertheless, in some extraordinary circumstances, the use of (12) might lead to a negative outcome of (11) as well. Sufficient conditions that the use of (12) leads to a nonnegative variance estimator with probability 1 are available from the authors upon request. For a general review of variance estimation methods in business surveys, see Brodie (2003).

Applying (12), a special problem may arise when $n_{h\ell}^t = 1$ or $n_{h\ell}^{t-12} = 1$. In order to evaluate the required sample variances, one may borrow the sample variance from a related substratum or from the same substratum in an earlier month. Alternatively, one may impute a variance when it emerges from the data that there is a relationship of the form $S_{h\ell}^2 \approx \sigma^2 \bar{O}_{h\ell}^2$; see Särndal, Swensson and Wretman (1992, page 461). In addition, the corresponding covariance term might be ignored when its (expected) contribution to the total variance is small. This is often the case when the sampling fractions in strata h and ℓ are small, that is in strata with relatively small units and, consequently, with small variances compared to the strata with larger units. Similar remarks apply to the imputed $\rho_{h\ell}^{t-12,t}$ when $n_{h\ell}^{t-12,t} \leq 2$ and $n_{h\ell}^{t-m} \geq 2$ ($m = 0, 12$). Since the $\rho_{h\ell}^{t-12,t}$ are often fairly high, this seems to be a viable way. In the example given in section 4 the $\rho_{h\ell}^{t-12,t}$ have an overall mean of 0.90 and a variance of 0.0074 so that the impact of the imputed $\rho_{h\ell}^{t-12,t}$ on the final results is likely to be moderate.

Furthermore, note that when $n_{h\ell}^{t-m} = 0$ ($m = 0$ or $m = 12$), the corresponding covariance term in (13) can be neglected without affecting its unbiasedness, provided that the remaining $S_{h\ell}^{t-12,t}$ are estimated in an unbiased way. Under this assumption such a term with $n_{h\ell}^{t-m} = 0$ ($m = 0$ or $m = 12$) can be neglected because the expectation of

$$n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}} \right) \hat{S}_{h\ell}^{t-12,t} \tag{14}$$

from (13) is equal to

$$\begin{aligned} E \left[E \left\{ n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}} \right) \hat{S}_{h\ell}^{t-12,t} \mid \mathbf{v}_{h\ell} \right\} \right] = \\ E \left\{ n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}} \right) S_{h\ell}^{t-12,t} \right\}, \end{aligned}$$

and the expectation on the right-hand side is the parameter to be estimated. Moreover, when $n_{h\ell}^{t-m} = 0$ ($m = 0$ or $m = 12$) and consequently $n_{h\ell}^{t-12,t} = 0$, the outcome of (14) is zero and the estimator $\hat{S}_{h\ell}^{t-12,t}$ for $S_{h\ell}^{t-12,t}$ becomes irrelevant. Therefore, ignoring such a term when $n_{h\ell}^{t-m} = 0$ ($m = 0$ or $m = 12$) does not affect the expectations of (13) and (14).

3.4 A comparison with Nordberg’s results

Using the standard formalism of inclusion indicators δ'_{hi} for each stratum, Nordberg (2000) derives a different expression for the first component in (6). However, it can be shown after some algebra that our expression (9) is equivalent to Nordberg’s (3.4); a proof is available from the authors upon request. In addition, Nordberg derives a non-zero expression for the second component in (6), i.e., the covariance between the two corresponding conditional expectations. Note that the Swedish sampling design is somewhat different from ours.

According to Nordberg (2000, page 370) the estimation of the second component for the Swedish sampling design requires a computer-intensive procedure which includes simulation of the sampling mechanism. However, since all $n'_{h\ell}$, $n_{h\ell}^{t-12}$ and $n_{h\ell}^{t-12,t}$ are ancillary statistics, an alternative might be to condition on these statistics so that the second component can be ignored. Recall that a statistic is called ancillary when its marginal distribution doesn’t depend on the target parameters to be estimated; see Cox and Hinkley (1974, pages 31-35). Such an alternative approach without the second component is to be recommended especially when $\hat{g}_{\text{STSRs}}^{t-12,t} \approx \hat{g}_{\text{PS,sub}}^{t-12,t}$ where $\hat{g}_{\text{PS,sub}}^{t-12,t}$ is the poststratified estimator based on the substrata $h\ell$. However, when the difference between $\hat{g}_{\text{STSRs}}^{t-12,t}$ and $\hat{g}_{\text{PS,sub}}^{t-12,t}$ is non-negligible, the calculation of the unconditional variance seems to be indispensable, including the estimation of the second component according to Nordberg. For a different approach to the estimation problem of the second component, see Appendix A.

For a justification of the use of a conditional (co)variance, see Holt and Smith (1979). An important advantage of the conditional (co)variance is that the corresponding confidence interval has better coverage properties than the one based on the unconditional variance. Denote the standard conditional 95% confidence interval for an arbitrary parameter θ by $(\hat{\theta}, \hat{\theta}_u | \mathbf{v})$ where \mathbf{v} denotes the vector consisting of all (ancillary) statistics involved in the conditional (co)variances. Then under the normality assumption and some mild conditions it holds that the actual 95% confidence level (CL) equals the nominal confidence level because

$$\begin{aligned} CL &= \sum_{v \in \Omega_v} P(v) P(\hat{\theta}_l < \theta < \hat{\theta}_u | v) \\ &= 0.95 \sum_{v \in \Omega_v} P(v) = 0.95, \end{aligned}$$

where Ω_v stands for the set of all possible outcomes of the random vector v . When unconditional (co)variances are used, the confidence intervals thus obtained may be quite inaccurate for a given sample allocation. Moreover, when averaged over all allocations CL may differ from 0.95; for an example, see Knottnerus (2003, pages 133-135). Note that in the planning stage before the sample is drawn, unconditional variances are always useful for examining Kish's design effect for a comparison of different sampling designs. In addition, note that for evaluating a conditional confidence interval for $g^{t-12,t}$ the underlying variances of $\hat{O}_{PS,sub}^{t-m}$ should also be taken conditional on the v_{ht} ($m = 0, 12$).

Finally, the unbiased estimator proposed by Nordberg [2000, Equation (3.9)] for the first component in (6) is quite different than those described in the previous subsection. In fact, his estimator is based on the following procedure for estimating the covariance term $S_{ht}^{t-12,t}$. Firstly, estimate the underlying quantity $\sum_{l=1}^{N_{ht}^{t-12,t}} O_{htl}^{t-12} O_{htl}^t$ from the overlap $s_{ht}^{t-12,t}$. Secondly, estimate the corresponding turnover means from S_{ht}^{t-12} and s_{ht}^t , respectively. Since the components thus estimated stem from different samples, a negative outcome of (11) cannot always be avoided. For a small example with real data, see the following section. In the remainder Nordberg's underlying estimator for $S_{ht}^{t-12,t}$ is denoted by $\hat{S}_{ht(NBG)}^{t-12,t}$. A derivation of the explicit expression for $\hat{S}_{ht(NBG)}^{t-12,t}$ is available from the authors upon request.

4. An application to the change of turnover in Dutch Supermarkets

4.1 Two estimators for the yearly change of turnover

For the impact on the variance estimators it is important to know that in January the turnover is estimated twice. The first estimate, denoted by \hat{O}^{janO} (with O for *old*), is made before the yearly sample update and is used to estimate the monthly change of the turnover in January compared to that in December. The second estimate, denoted by \hat{O}^{janN} (with N for *new*), is made after the yearly sample update and is used to estimate the monthly change of the turnover in February compared to January. This procedure implies that units of the old sample as well as those of the new sample receive a questionnaire in January.

Unlike estimator (1) the actual estimator used by Statistics Netherlands for the yearly change in the monthly turnover is based on a chain of 12 monthly changes in turnover

$$\begin{aligned} \hat{G}_{act}^{t,t-12} &= 1 + \hat{g}_{act}^{t,t-12} = \prod_{j=0}^{11} (1 + \hat{g}^{t, t-j-1}) \\ &= \frac{\hat{O}^t}{\hat{O}^{t-1}} \times \frac{\hat{O}^{t-1}}{\hat{O}^{t-2}} \times \dots \times \frac{\hat{O}^{feb}}{\hat{O}^{janN}} \\ &\quad \times \frac{\hat{O}^{janO}}{\hat{O}^{dec}} \times \dots \times \frac{\hat{O}^{t-11}}{\hat{O}^{t-12}} \\ &= \frac{\hat{O}^t}{\hat{O}^{t-12}} \times \frac{\hat{O}^{janO}}{\hat{O}^{janN}} \quad (t \neq \text{jan}). \end{aligned} \quad (15)$$

In this section we will compare the variances of estimators (1) and (15). Similar to (2) the variance formulas for $\hat{g}_{act}^{t,t-12}$ can be derived by a first-order Taylor series expansion.

4.2 Description of the data

The calculations for the variances and confidence intervals in this example are based on turnover data of Dutch Supermarkets of 4-week periods in 2003 and 2004 (*i.e.*, $t = 1, \dots, 26$). Hence, there are 13 observations in one year and, consequently, we use slightly adjusted symbols such as $g^{t,t-13}$ in the remainder of this section.

The population consists of about 3,500 establishments. The turnover data stem from a stratified sample and administrative files. A gross STSRS sample of about 900 units stratified by size is drawn from the full list of population units of the GBR that includes the units of the administrative files as well. Establishments with 50 or more *employees* are included with probability 1. The other establishments are sampled with decreasing inclusion probability from 1:2 (20-49 employees per establishment) to 1:40 in the smallest size (1 employee per establishment). The administrative files contain about 950 units, present in all size classes. About 500 of the 900 units in the gross sample were already present in the administrative files, but they do not receive a questionnaire. Thus, the net sample contains about 400 units. In fact, the sample size for each stratum in this specific example is random. However, as explained in subsection 3.4, we estimate all (co)variances conditional on the n_h in such a case. Data from units within the administrative files are put into a separate stratum with the sampling fraction being unity.

4.3 Results

Table 1 gives the yearly growth rates and their 95% margins for $t = 16, \dots, 24$. It emerges that the 95% margins for the estimated growth rates $\hat{g}_{act}^{t,t-13}$, currently used by Statistics Netherlands, vary between 0.8 and 1.0 (per cent point). For example, in the first period ($t = 16$) the 95% confidence interval for the yearly growth rate is -1.3 to 0.7 per cent. As expected, the 95% margins for the more complicated estimator $\hat{g}_{act}^{t,t-13}$ are close to those for the simpler $\hat{g}^{t,t-13}$ from (1). The 95% margins of $\hat{g}^{t,t-13}$ vary

between 0.7 and 1.0 (per cent point). The estimator for the growth rate to be preferred is $\hat{g}_{act}^{t,t-13}$ as it corrects for the yearly sample update in January. The estimation of its variance, however, can be simplified by using the variance estimator described in section 3 rather than the more laborious expression for $\text{var}(\hat{g}_{act}^{t,t-13})$.

Table 1
Estimated growth rates with 95% margins

t	$\hat{g}_{act}^{t,t-13} \times 100\%$	$\hat{g}^{t,t-13} \times 100\%$
16	-0.3 (± 1.0) ¹	-0.4 (± 1.0)
17	-3.7 (± 1.0)	-3.8 (± 0.9)
18	1.6 (± 1.0)	1.5 (± 0.9)
19	-2.2 (± 0.9)	-2.3 (± 0.9)
20	0.5 (± 0.8)	0.4 (± 0.7)
21	-1.7 (± 0.8)	-1.8 (± 0.7)
22	-2.2 (± 0.8)	-2.3 (± 0.7)
23	0.0 (± 0.8)	-0.1 (± 0.7)
24	-2.3 (± 0.9)	-2.4 (± 0.9)

¹ The 95% margins are given between parentheses.

As described in section 3, we have used the estimated correlation $\hat{\rho}_{h\ell,OLP}^{t-13,t}$ from the overlap $s_{h\ell}^{t-13,t}$ to estimate covariance $S_{h\ell}^{t-13,t}$ in order to avoid negative outcomes of (11). Knottnerus and Van Delden (2006) evaluated the bias of $\hat{\rho}_{h\ell,OLP}^{t-13,t}$ for the Dutch Supermarket data and found a small underestimation of $\hat{\rho}_{h\ell,OLP}^{t-13,t}$ resulting in a minor, less than 5%, overestimation of $\text{var}(\hat{g}^{t,t-13})$.

The use of estimator $\hat{S}_{h\ell,OLP}^{t-13,t}$ in (10) may give a negative outcome of (11) and an estimated correlation $\hat{\rho}_{h\ell}^{t-13,t}$ larger than 1. For example, consider the specific population with $N = 50$ and $H = 1$ consisting of the units of substratum $h\ell = 65$. From the panel data for this population, given in Table 2 for $t = 3$ and $t = 16$, we obtain after some calculations $\hat{S}^{t-13} = 410.7$, $\hat{S}^t = 394.3$ and $\hat{G}^{t,t-13} = 1.028$. Note that in the remainder of this section the subscript $h\ell = 11$ is omitted in the symbols because there is only one stratum. Table 3 gives, for three different approaches, some additional estimates for the panel data in Table 2. For example, using $\hat{S}_{OLP}^{t-13,t}$ in (10) results in an estimated

correlation $\hat{\rho}^{t-13,t} = 1.39$. This then yields a negative variance estimate from (11) of minus 2.2 million. Likewise, for the same data the alternative estimator $\hat{S}_{NBG}^{t-13,t}$ of $S^{t-13,t}$ based on Nordberg (2000) results in minus 36.9 million as outcome of (11) because the corresponding estimate $\hat{\rho}_{NBG}^{t-13,t}$ becomes 1.64. In contrast, using the correlation estimated from the overlapping sample $s^{t-13,t}$ according to (12) yields $\hat{\rho}_{OLP}^{t-13,t} = 0.9997$ and the positive variance estimate from (11) becomes 52.1 million. In addition, for the panel data in Table 2 the outcome of Nordberg's estimator (3.9) for the covariance between \hat{O}^{t-13} and \hat{O}^t is 111.1 million whereas covariance estimator (13) proposed here yields 67.8 million.

Table 2
Panel data¹ from a population with $N = 50$ and $H = 1$

period	turnover per unit (in thousand euros)				
	1	2	3	4	5
$t = 3$	493.9	264.3	1,179.1	380.0	
$t = 16$	475.3	472.0	267.0	1,169.0	

¹ Actually, the panel data belonged to substratum $h\ell = 65$.

5. Conclusions

The variance formulas obtained in this paper are useful for calculating the variance of an estimated yearly growth rate of monthly turnover. The use of (13) as an estimator for $\text{cov}(\hat{O}^{t-12}, \hat{O}^t)$ results in reasonable estimates of the covariance of change in particular. The variance estimation procedure allows for rotating panels, births, deaths, and units that migrate between strata.

Furthermore, we recommend estimating a population covariance according to (12) based on the corresponding correlation estimated from the overlap and on the corresponding variances estimated from the larger separate samples. This may help to avoid a serious underestimation or a negative outcome of the variance estimator for the yearly growth rate. The resulting estimated covariances are only slightly biased.

Table 3
Estimates from three different approaches

approach		parameters to be estimated		
		$S^{t-13,t}$	$\rho^{t-13,t}$	$\text{var}(\hat{O}^t - G^{t,t-13}\hat{O}^{t-13})$
Nordberg(2000)	estimator	$\hat{S}_{NBG}^{t-13,t}$	$\frac{\hat{S}_{NBG}^{t-13,t}}{\hat{S}^{t-13}\hat{S}^t}$	Eq. (11)
	result	265.2×10^3	1.64	-36.9×10^6
Eq. (10)	estimator	$\hat{S}_{OLP}^{t-13,t}$	$\frac{\hat{S}_{OLP}^{t-13,t}}{\hat{S}^{t-13}\hat{S}^t}$	Eq. (11)
	result	225.0×10^3	1.39	-2.2×10^6
Eq. (12)	estimator	$\hat{\rho}_{OLP}^{t-13,t} \hat{S}^{t-13}\hat{S}^t$	$\hat{\rho}_{OLP}^{t-13,t}$	Eq. (11)
	result	161.9×10^3	1.00 ¹	52.1×10^6

¹ In fact, 0.9997.

For the sampling design of the Dutch Supermarkets the second covariance term in (6) is negligible due to the fact that $n_{h\ell}^{\text{dec, jan}}_{\text{REQ}}$ is fixed. In contrast, for the SAMU design in Sweden this term is non-negligible and its estimation is time-consuming; the word SAMU (SAMordnade Urval) is a Swedish acronym for coordinated samples. In Appendix A we propose an alternative method for estimating this covariance. However, under the condition that $\hat{g}^{t,t-12} \approx \hat{g}_{\text{PS, sub}}^{t,t-12}$ it suffices in our opinion to only use the first covariance. This simplifies the estimation procedure considerably. Moreover, under the normality assumption the conditional confidence interval has better coverage properties compared to the unconditional interval.

The example of the Dutch Supermarkets shows one of the practical applications of the variance formulas: determining which estimator has the smallest variance. The results confirm that the variance of the simple estimator $\hat{g}^{t,t-13}$ is close to that of $\hat{g}_{\text{act}}^{t,t-13}$ from section 4 which corrects for the sample refreshment in January. Hence, for the Dutch Supermarkets $\text{var}(\hat{g}^{t,t-13})$ might be used for estimating $\text{var}(\hat{g}_{\text{act}}^{t,t-13})$. For branches with another SIC code it needs to be checked whether $\text{var}(\hat{g}_{\text{act}}^{t,t-13}) \approx \text{var}(\hat{g}^{t,t-13})$ since the impact of the refreshment in January need not be negligible.

Acknowledgements

The views expressed in the paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors would like to thank the Associate Editor and two anonymous referees for their useful comments and suggestions, which have led to a significant improvement of this paper.

Appendix A

Justification of (5)

Firstly, consider the case of strata without births and deaths. Apart from the yearly update in January, there are now no monthly updates. Hence, $n_{h\ell}^t = n_{h\ell}^{\text{dec, jan}}_{\text{REQ}}$ is fixed from which (5) follows. This case applies to the Dutch Supermarkets because that population has been quite stable over the years. Secondly, in case of births and deaths among the strata we can write $n_{h\ell}^t$ as

$$n_{h\ell}^t = n_{\ell}^t - n_{0\ell}^{t-12,t} - \sum_{k \neq h} n_{k\ell}^t, \quad (16)$$

where $n_{0\ell}^{t-12,t}$ or, for short, $n_{0\ell}^t$ stands for the number of births in months $t-12, \dots, t-1$ among s_{ℓ}^t . Because the sampling procedure among the new births after month $t-12$ is independent of the $n_{h\ell}^{t-12,t}$, the random variables $n_{0\ell}^{t-12,t}$ and $n_{h\ell}^{t-12,t}$ have a zero covariance. Furthermore, using

$\text{cov}(n_{h\ell}^{t-12,t}, n_{k\ell}^t) = 0$ for $k \neq h$, it is seen from (16) that $\text{cov}(n_{h\ell}^{t-12,t}, n_{h\ell}^t) = 0$ ($h = 1, \dots, H$).

In fact, it is assumed so far that the distribution of $n_{k\ell}^t$ ($k \neq h$) can be described by a hypergeometric distribution with parameters $(N_{\ell}^t, N_{k\ell}^{t-12,t}, n_{\ell}^t)$ irrespective of the values of the $n_{h\ell}^{t-12,t}$. A similar remark applies to $n_{0\ell}^{t-12,t}$. However, it can be argued that in practice these assumptions lead to a minor, second-order error in the variance formulas. In order to trace this error, we assume for simplicity's sake and without loss of generality that (i) births and deaths do not migrate between strata, (ii) there are no deaths among the births, (iii) $n_{0h}^t = f_h N_{0h}^t$ is fixed, (iv) after their first month in the population births are irrelevant for the monthly updates during the rest of the study period and (v) deaths are not selected in or removed from the sample by the monthly updates; so a third-order error is still ignored. Under these assumptions we now look more closely at the second covariance component for $\ell = h$, say $C_{hh, \text{sec}}$, from (4a). In analogy with (6) $C_{hh, \text{sec}}$ can be written as

$$\begin{aligned} C_{hh, \text{sec}} &= \frac{1}{n_h^{t-12} n_h^t} \\ &\text{cov} \left\{ E \left(\sum_{g=1}^{H+1} n_{hg}^{t-12} \bar{o}_{hg}^{t-12} |v_h \right), E \left(\sum_{k=0}^H n_{kh}^t \bar{o}_{kh}^t |v_h \right) \right\} \\ &= \frac{1}{n_h^{t-12} n_h^t} \sum_{g=1}^{H+1} \sum_{k=1}^H \bar{o}_{hg}^{t-12} \bar{o}_{kh}^t \text{cov}(n_{hg}^{t-12}, n_{kh}^t), \end{aligned} \quad (17)$$

where $v_h = (n_{h1}^{t-12}, \dots, n_{h, H+1}^{t-12}, n_{1h}^t, \dots, n_{Hh}^t)$. Note that under the above assumptions $C_{h\ell, \text{sec}} = 0$ for $\ell \neq h$.

To estimate the covariances in (17), consider the formula for the conditional expectation of y given $x = x_0$ when y and x follow a bivariate normal distribution. That is, in standard notation,

$$E(y|x_0) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2} (x_0 - \mu_x).$$

In addition, for a given change Δx_0 of x the conditional expectation of the change of y is equal to $E(\Delta y | \Delta x_0) = \sigma_{yx} \Delta x_0 / \sigma_x^2$ or, equivalently,

$$\sigma_{yx} = \frac{E(\Delta y | \Delta x_0)}{\Delta x_0} \sigma_x^2. \quad (18)$$

So for estimating, for instance, $\text{cov}(n_{h, H+1}^{t-12}, n_{kh}^t)$ in (17) under normality it suffices to evaluate the expected effect on $y = n_{kh}^t$ caused by a change of the future deaths $x = n_{h, H+1}^{t-12}$ in s_h^{t-12} .

Let $\Delta n_{h, H+1}^{t-12}$ denote an additional (positive) change of these deaths in s_h^{t-12} . Define $p_{h, H+1}^{\text{jan}, t}$ by $p_{h, H+1}^{\text{jan}, t} = N_{h, H+1}^{\text{jan}, t} / N_{h, H+1}^{t-12}$ where $N_{h, H+1}^{\text{jan}, t}$ is the number of deaths in stratum h between January and month t . Also, $p_{hg}^{t-12} = N_{hg}^{t-12,t} / N_h^{t-12}$ ($g = 1, \dots, H+1$). Using assumption (v), the expected number of additional deaths in the sample of January before

the refreshment can be estimated by $p_{h,H+1}^{\text{jan},t} \Delta n_{h,H+1}^{t-12}$. Subsequently, the expected number of additional deaths in the sample after the refreshment can be estimated by

$$\gamma_{\text{red}}^{\text{jan}} p_{h,H+1}^{\text{jan},t} \Delta n_{h,H+1}^{t-12};$$

$$\gamma_{\text{red}}^{\text{jan}} = (0.9 - f_h) / (1 - f_h), \quad (19)$$

where $\gamma_{\text{red}}^{\text{jan}}$ is the reduction factor due to the refreshment in January. For the derivation of (19), see the end of this appendix. The corresponding monthly updates between January and month t due to these additional deaths in the sample from stratum h lead to the following estimate of the expected increase of *incoming* units n_{kh}^t from stratum k ($k \neq h$) in the sample of month t

$$E(\Delta n_{kh}^t | \Delta n_{h,H+1}^{t-12}) = \gamma_{\text{red}}^{\text{jan}} p_{h,H+1}^{\text{jan},t} \Delta n_{h,H+1}^{t-12} p_{kh}^t, \quad (20)$$

where $p_{kh}^t = N_{kh}^{t-12,t} / (N_h^t - N_{0h}^t)$. Recall from subsection 2.1 that an update in month s occurs only when $d_h^{s-1} \neq f_h D_h^{s-1}$, where D_h^s (d_h^s) stands for the number of deaths in U_h^s (S_h^s), and that $n_{kh}^t = f_h N_{kh}^{t-12,t}$ is fixed when $N_{h,H+1}^{\text{jan},t} = 0$ ($k \neq h$). Furthermore, note that births are excluded in the definition of p_{kh}^t in (20) because of assumption (iv).

Next, define for $m = 0, 12$

$$\bar{O}_h^{t-m} = \frac{1}{N_h^{t-m}} \sum_{i=1}^{N_h^{t-m}} O_{hi}^{t-m};$$

$$(S_h^{t-m})^2 = \frac{1}{N_h^{t-m} - 1} \sum_{i=1}^{N_h^{t-m}} (O_{hi}^{t-m} - \bar{O}_h^{t-m})^2;$$

$$p_{h,\leq H}^{t-12} = 1 - p_{h,H+1}^{t-12};$$

$$p_{in,h}^t = 1 - p_{hh}^t;$$

$$\bar{O}_{h,\leq H}^{t-12} = \sum_{g=1}^H \frac{p_{hg}^{t-12}}{p_{h,\leq H}^{t-12}} \bar{O}_{hg}^{t-12};$$

$$\bar{O}_{in,h}^t = \sum_{\substack{k=1 \\ k \neq h}}^H \frac{p_{kh}^t}{p_{in,h}^t} \bar{O}_{kh}^t.$$

Now using (18) and (20), we obtain for $k \neq h$ the following covariance approximation

$$\text{acov}(n_{h,H+1}^{t-12}, n_{kh}^t)$$

$$= \frac{E(\Delta n_{kh}^t | \Delta n_{h,H+1}^{t-12})}{\Delta n_{h,H+1}^{t-12}} \text{var}(n_{h,H+1}^{t-12})$$

$$\approx \gamma_{\text{red}}^{\text{jan}} p_{h,H+1}^{\text{jan},t} p_{kh}^t n_{h,H+1}^{t-12} p_{h,H+1}^{t-12} (1 - p_{h,H+1}^{t-12}) (1 - f_h)$$

$$= n_h^{t-12} p_{kh}^t A_h / p_{in,h}^t; \quad (21)$$

$$A_h = \gamma_{\text{red}}^{\text{jan}} p_{in,h}^t p_{h,H+1}^{\text{jan},t} p_{h,H+1}^{t-12} (1 - p_{h,H+1}^{t-12}) (1 - f_h),$$

where, for simplicity, we omitted the term $N_h^{t-12} / (N_h^{t-12} - 1)$ in the second line. Because n_h^{t-12} is fixed, it holds that $\text{cov}(n_{h,H+1}^{t-12}, n_{kh}^t) = -\text{cov}(n_{h1}^{t-12} + \dots + n_{hH}^{t-12}, n_{kh}^t)$. Hence, in analogy with the multihypergeometric distribution

we can use for $1 \leq g \leq H$ and $k \neq h$ the following relationship for an approximation of $\text{cov}(n_{hg}^{t-12}, n_{kh}^t)$

$$\text{acov}(n_{hg}^{t-12}, n_{kh}^t) = -\frac{p_{hg}^{t-12}}{p_{h,\leq H}^{t-12}} \text{acov}(n_{h,H+1}^{t-12}, n_{kh}^t) \quad (22a)$$

$$= -n_h^{t-12} \frac{p_{hg}^{t-12}}{p_{h,\leq H}^{t-12}} \frac{p_{kh}^t}{p_{in,h}^t} A_h, \quad (22b)$$

where (21) is used as well. Alternatively, note that

$$\text{cov}(n_{h,H+1}^{t-12}, n_{kh}^t) = -\text{cov}(n_{h,\leq H}^{t-12}, n_{kh}^t)$$

$$= -\sum_{g < H} \sum_{i \in U_{hg}^{t-12,t}} \text{cov}(\delta_{hgi}^{t-12}, n_{kh}^t),$$

where

$$\delta_{hgi}^{t-12} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ unit in } U_{hg}^{t-12,t} \text{ is included in sample } s_h^{t-12} \\ 0 & \text{otherwise.} \end{cases}$$

Hence, by symmetry, $\text{cov}(\delta_{hgi}^{t-12}, n_{kh}^t) = -\text{cov}(n_{h,H+1}^{t-12}, n_{kh}^t) / N_{h,\leq H}^{t-12,t}$ from which (22a) follows ($1 \leq g \leq H$). Likewise, for $k = h$ we obtain from (21) and (22b)

$$\text{acov}(n_{h,H+1}^{t-12}, n_{hh}^t) = -n_h^{t-12} A_h; \quad (23)$$

$$\text{acov}(n_{hg}^{t-12}, n_{hh}^t) = n_h^{t-12} p_{hg}^{t-12} A_h / p_{h,\leq H}^{t-12},$$

respectively ($1 \leq g \leq H$). Now substituting (21)–(23) into (17), we get the approximation

$$C_{hh,\text{sec}} = A_h (\bar{O}_{h,H+1}^{t-12} - \bar{O}_{h,\leq H}^{t-12}) (\bar{O}_{in,h}^t - \bar{O}_{hh}^t) / n_h^t. \quad (24)$$

Assuming that the two terms between parentheses in (24) are absolutely smaller than S_h^t , it follows from (24) that

$$|C_{hh,\text{sec}}| \leq \frac{\gamma_{\text{red}}^{\text{jan}} p_{h,H+1}^{\text{jan},t} p_{in,h}^t p_{h,H+1}^{t-12} (1 - p_{h,H+1}^{t-12}) (1 - f_h)}{n_h^t} (S_h^t)^2.$$

Hence, when $p_{in,h}^t, p_{h,H+1}^{t-12} \leq 0.1$, we may conclude that under the above assumptions the contribution of the second covariance component is less than 1% of $\text{var}(\bar{O}_h^t)$ so that (5) can be used without severely affecting the results. When $C_{hh,\text{sec}}$ is non-negligible, it can be estimated from the sample according to (24) by

$$\hat{C}_{hh,\text{sec}} = A_h \{(\bar{O}_{h,H+1}^{t-12} - \bar{O}_{h,\leq H}^{t-12})(\bar{O}_{in,h}^t - \bar{O}_{hh}^t) - \text{c}\hat{\text{ov}}(\bar{O}_{h,\leq H}^{t-12}, \bar{O}_{hh}^t)\} / n_h^t, \quad (25)$$

where in analogy with (10) and (12) $\text{c}\hat{\text{ov}}(\bar{O}_{h,\leq H}^{t-12}, \bar{O}_{hh}^t)$ is defined by

$$\text{c}\hat{\text{ov}}(\bar{O}_{h,\leq H}^{t-12}, \bar{O}_{hh}^t) = \frac{n_{hh}^{t-12}}{n_{h,\leq H}^{t-12}} \left(\frac{n_{hh}^{t-12,t}}{n_{hh}^{t-12} n_{hh}^t} - \frac{1}{N_h^{t-12,t}} \right) \hat{\rho}_{hh,\text{OLP}}^{t-12,t} \hat{S}_{hh}^t \hat{S}_{hh}^t.$$

We used in (25) that (i) for two arbitrary (unbiased) estimators \hat{a} and \hat{b} , $E(\hat{a}\hat{b}) = ab + \text{cov}(\hat{a}, \hat{b})$ and (ii) $\text{cov}(\bar{O}_{hg}^{t-12}, \bar{O}_{kh}^t) = 0$ ($g \neq h$ or $k \neq h$).

We conclude this appendix with the derivation of (19). The expected number of additional deaths remaining in the sample of January during the refreshment is $0.9 p_{h,H+1}^{\text{jan},t} \Delta n_{h,H+1}^{t-12}$. The number of deaths outside the sample just

before the refreshment can be estimated by $N_{h,H+1}^{\text{jan},t} - n_{h,H+1}^{\text{jan}} - p_{h,H+1}^{\text{jan},t} \Delta n_{h,H+1}^{t-12}$. Hence, the number of new deaths in the sample due to the refreshments in all substrata $U_{hg}^{t-12,t}$ ($1 \leq g \leq H$) in January can be estimated by

$$0.1(n_h^{\text{jan}} - n_{0h}^{\text{jan}}) \frac{N_{h,H+1}^{\text{jan},t} - n_{h,H+1}^{\text{jan}} - p_{h,H+1}^{\text{jan},t} \Delta n_{h,H+1}^{t-12}}{N_h^{\text{jan}} - N_{0h}^{t-12,\text{jan}} - (n_h^{\text{jan}} - n_{0h}^{\text{jan}})}.$$

Now using $n_{0h}^{\text{jan}} = f_h N_{0h}^{t-12,\text{jan}}$ according to the above assumptions, it is seen that after the refreshments the final number of additional deaths in the sample due to $\Delta n_{h,H+1}^{t-12,t}$ can be estimated by

$$\begin{aligned} & \left\{ 0.9 - \frac{0.1(n_h^{\text{jan}} - n_{0h}^{\text{jan}})}{N_h^{\text{jan}} - N_{0h}^{t-12,\text{jan}} - (n_h^{\text{jan}} - n_{0h}^{\text{jan}})} \right\} p_{h,H+1}^{\text{jan},t} \Delta n_{h,H+1}^{t-12} \\ &= \frac{0.9 - f_h}{1 - f_h} p_{h,H+1}^{\text{jan},t} \Delta n_{h,H+1}^{t-12} = \gamma_{\text{red}} p_{h,H+1}^{\text{jan},t} \Delta n_{h,H+1}^{t-12}. \end{aligned}$$

Appendix B

Some useful covariance formulas for overlapping samples

Let s_{123} denote a mother sample consisting of three mutually disjoint SRS subsamples s_1, s_2 and s_3 . Let the variable x be observed in s_{12} and the variable y in s_{23} . The corresponding sample means are denoted by \bar{x}_{12} and \bar{y}_{23} , respectively. Denote the size of s_k by n_k ($k = 1, 2, 3, 12, 23$). Define $\lambda = n_2/n_{12}$, $\mu = n_2/n_{23}$ and $f_k = n_k/N$. Furthermore, define S_{xy} by

$$S_{xy} = \frac{1}{N-1} \sum_{j=1}^N (X_j - \bar{X}_p)(Y_j - \bar{Y}_p).$$

Then the covariance between \bar{x}_{12} and \bar{y}_{23} is equal to

$$\text{cov}(\bar{x}_{12}, \bar{y}_{23}) = \left(\frac{\lambda\mu}{n_2} - \frac{1}{N} \right) S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N} \right) S_{xy} = \left(\frac{\lambda}{n_{23}} - \frac{1}{N} \right) S_{xy}. \quad (26)$$

This can be shown as follows

$$\begin{aligned} \text{cov}(\bar{x}_{12}, \bar{y}_{23}) &= \text{cov}\{(1-\lambda)\bar{x}_1 + \lambda\bar{x}_2, \mu\bar{y}_2 + (1-\mu)\bar{y}_3\} \\ &= (1-\lambda)\text{cov}(\bar{x}_1, \bar{y}_{23}) + \lambda\mu\text{cov}(\bar{x}_2, \bar{y}_2) \\ &\quad + \lambda(1-\mu)\text{cov}(\bar{x}_2, \bar{y}_3) \\ &= -(1-\lambda)\frac{S_{xy}}{N} + \lambda\mu\left(\frac{1}{n_2} - \frac{1}{N}\right)S_{xy} - \lambda(1-\mu)\frac{S_{xy}}{N} \\ &= \left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right)S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N}\right)S_{xy} = \left(\frac{\lambda}{n_{23}} - \frac{1}{N}\right)S_{xy}. \end{aligned}$$

In the third line we used that $\text{cov}(\bar{x}_1, \bar{y}_{23}) = \text{cov}(\bar{x}_2, \bar{y}_3) = -S_{xy}/N$. This follows from the conditional covariance formula

$$\begin{aligned} \text{cov}(\bar{x}_2, \bar{y}_3) &= E\{\text{cov}(\bar{x}_2, \bar{y}_3 | s_2)\} + \text{cov}\{E(\bar{x}_2 | s_2), E(\bar{y}_3 | s_2)\} \\ &= 0 + \text{cov}\left\{\bar{x}_2, \frac{\bar{Y}_p - f_2\bar{y}_2}{1 - f_2}\right\} \\ &= -\frac{f_2}{1 - f_2} \text{cov}(\bar{x}_2, \bar{y}_2) = -\frac{S_{xy}}{N}. \end{aligned}$$

For an alternative proof based on the sampling autocorrelation coefficient, see Knottnerus (2003, page 375).

References

- Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics*, 32, 451-467.
- Brodie, P. (2003). Review of recent work on variance estimation methods in business surveys. Unpublished report, Office for National Statistics, London.
- Cox, D.R., and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Hidioglou, M.A., Särndal, C.-E. and Binder, D.A. (1995). Weighting and estimation in business surveys. In *Business Survey Methods*, (Eds., B.G. Cox et al.). New York: John Wiley & Sons, Inc.
- Holt, D., and Skinner, C.J. (1989). Components of change in repeated surveys. *International Statistical Review*, 57, 1-18.
- Holt, D., and Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, A, 142, 33-46.
- Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons, Inc.
- Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. New York: Springer-Verlag.
- Knottnerus, P., and Van Delden, A. (2006). Estimation of changes in repeated surveys and their significance. <http://www.iser.essex.ac.uk/ulsc/mol2006/programme/data/paper/Knottnerus.doc>.
- Konschnik, C.A., Monsour, N.J. and Detlefsen, R.E. (1985). Constructing and maintaining frames and samples for business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 113-122.
- Laniel, N. (1987). Variances for a rotating sample from a changing population. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 496-500.
- Lowerre, J.M. (1979). Sampling for change. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 343-347.
- Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.
- Qualité, L., and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34, 173-181.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Smith, P., Pont, M. and Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994-2000. *The Statistician*, 52, 257-295.
- Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, 288-289.
- Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, 24, 53-78.

Variance inflation factors in the analysis of complex survey data

Dan Liao and Richard Valliant¹

Abstract

Survey data are often used to fit linear regression models. The values of covariates used in modeling are not controlled as they might be in an experiment. Thus, collinearity among the covariates is an inevitable problem in the analysis of survey data. Although many books and articles have described the collinearity problem and proposed strategies to understand, assess and handle its presence, the survey literature has not provided appropriate diagnostic tools to evaluate its impact on regression estimation when the survey complexities are considered. We have developed variance inflation factors (VIFs) that measure the amount that variances of parameter estimators are increased due to having non-orthogonal predictors. The VIFs are appropriate for survey-weighted regression estimators and account for complex design features, *e.g.*, weights, clusters, and strata. Illustrations of these methods are given using a probability sample from a household survey of health and nutrition.

Key Words: Cluster sample; Collinearity diagnostics; Linearization variance estimator; Survey-weighted least squares; Stratified sample.

1. Introduction

Collinearity of predictor variables in a linear regression refers to a situation where explanatory variables are correlated with each other. The terms, multicollinearity and ill conditioning, are also used to denote the same situation. Collinearity is worrisome for both numerical and statistical reasons. The estimates of slope coefficients can be numerically unstable in some data sets in the sense that small changes in the \mathbf{X} 's or the \mathbf{Y} 's can produce large changes in the values of these estimates. Statistically, correlation among the predictors can lead to slope estimates with large variances. In addition, when \mathbf{X} 's are strongly correlated, the R^2 in a regression can be large while the individual slope estimates are not statistically significant. Even if slope estimates are significant, they may have signs that are the opposite of what are expected (Neter, Kutner, Wasserman and Nachtsheim 1996). Collinearity may also affect forecasts (Smith 1974; Belsley 1984).

In experimental designs, it may be possible to create situations where the explanatory variables are orthogonal to each other. But, in many surveys, variables that are substantially correlated are collected for analysis. For example, total income and its components (*e.g.*, wages and salaries, capital gains, interest and dividends) are collected in the Panel Survey of Income Dynamics (<http://psidonline.isr.umich.edu/>) to track economic well-being over time. When one explanatory variable is a linear combination of the others, this is known as perfect collinearity (or multicollinearity) and is easy to identify. Cases that are of interest in practice are ones where collinearity is less than perfect but still affects the precision of estimates (Kmenta 1986, section 10.3).

Although there is a substantial literature on regression diagnostics for non-survey data, there is considerably less for survey data. A few articles in the last decade introduced techniques for the evaluation of the quality of regression on complex survey data, mainly on identifying influential points and influential groups with abnormal data values or survey weights. Elliot (2007), for instance, developed Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. Li (2007a,b); Li and Valliant (2009, 2011) adapted and extended a series of traditional diagnostic techniques to regression on complex survey data, mainly on identifying influential observations and influential groups of observations. Li's research covers residuals and leverages, DFBETA, DFBETAS, DFFIT, DFFITS, Cook's Distance and the forward search approach. Although an extensive literature in applied statistics provides valuable suggestions and guidelines for data analysts to diagnose the presence of collinearity (*e.g.*, Farrar and Glauber 1967; Theil 1971; Belsley, Kuh and Welsch 1980; Fox 1984; Belsley 1991), none of this research touches upon diagnostics for collinearity when fitting models with survey data.

The variance inflation factor (VIF) described in section 2, is one of the most popular conventional collinearity diagnostic techniques, and is mainly aimed at ordinary or weighted least squares regressions. A VIF measures the inflation of the variance of a slope estimate caused by the nonorthogonality of the predictors over and above what the variance would be with orthogonality. In section 3, we consider the case of an analyst who estimates model parameters using survey-weighted least squares (SWLS) and derive VIFs appropriate to SWLS. The components of the VIF can be estimated using the ingredients of a variance

1. Dan Liao, RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005. E-mail: dliao@rti.org; Richard Valliant, University of Michigan and University of Maryland, Joint Program in Survey Methodology, 1218 Lefrak Hall, College Park, MD, 20742.

estimator that is in common usage in software packages for analyzing survey data. In the case of linear regression, a type of sandwich variance estimator will estimate both the model variance and design variance of the SWLS slope estimator. As we will show in section 3, the model or design variance of $\hat{\beta}_k$, an estimator of slope associated with the predictor \mathbf{x}_k , is inflated somewhat when different predictors are correlated with each other compared to what the variance would be if \mathbf{x}_k were orthogonal to the other predictors. The measure of inflation, the VIF, is composed of terms that must be estimated from the sample. Our approach has been to substitute estimators that have both a model and design interpretation as described in section 3.5.

The fourth section presents an empirical study using data from the United States National Health and Nutrition Examination Survey. The application of our new approach is demonstrated and the newly-derived VIF values for SWLS are compared to the ones for OLS or WLS, which can be obtained from the standard statistical packages. The comparisons show that VIF values are different for different regression methods and a VIF specific to complex sample should be used to evaluate the harmfulness of collinearity in the analysis of survey data.

2. Collinearity diagnostics in ordinary least squares estimation

Suppose the sample s has n units, on each of which p \mathbf{x} 's or predictors and one analysis variable Y are observed. The standard linear model in a nonsurvey setting is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where \mathbf{Y} is an $n \times 1$ vector of observations on a response or dependent variable; $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ is an $n \times p$ design matrix of fixed constants with \mathbf{x}_k , the $n \times 1$ vector of values of explanatory variable k for the n sample units; $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters to be estimated; and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of statistically independent error terms with zero mean and constant variance σ^2 . We assume, for simplicity, that \mathbf{X} has full column rank. The ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, for which the model variance is $\text{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$. Here, we use the subscript M to denote expectation under the model.

Collinearities of explanatory variables inflate the model variance of the regression coefficients compared to having orthogonal \mathbf{x} 's. This effect can be seen in the formula for the variance of a specific estimated non-intercept coefficient $\hat{\beta}_k$ (Theil 1971),

$$\text{Var}_M(\hat{\beta}_k) = \frac{\sigma^2}{\sum_{i \in s} x_{ik}^2} \frac{1}{1 - R_k^2} \quad (1)$$

where R_k^2 is the square of the multiple correlation from the regression of the k^{th} column of \mathbf{X} on the other columns. This R-square defined as $R_k^2 = \hat{\beta}_{(k)}^T \mathbf{X}_{(k)}^T \mathbf{X}_{(k)} \hat{\beta}_{(k)} / \mathbf{x}_k^T \mathbf{x}_k$, where $\hat{\beta}_{(k)}$ is OLS estimate of the slope when \mathbf{x}_k is regressed on the other \mathbf{x} 's and $\mathbf{X}_{(k)}$ is the \mathbf{X} matrix with the k^{th} column removed. The term $\sigma^2 / \sum x_{ik}^2$ is the model variance of $\hat{\beta}_k$ if the k^{th} predictor were orthogonal to all the other predictors. The value of R_k^2 may be nonzero because the k^{th} predictor is correlated with one other explanatory variable or because of a more complex pattern of dependence between \mathbf{x}_k and several other predictors. Consequently, the collinearity between \mathbf{x}_k and some other explanatory variables can result in the inflation of the variance of $\hat{\beta}_k$ beyond what would be obtained with orthogonal \mathbf{x} 's. The second term in (1), $(1 - R_k^2)^{-1}$, is called the variance-inflation factor (VIF) (Theil 1971).

A basic reference on collinearity and other OLS diagnostics is Belsley *et al.* (1980). Collinearity diagnostics are covered in many other textbooks including Fox (1984) and Neter *et al.* (1996). In some cases, it is desirable to weight cases differentially in a regression analysis to incorporate a nonconstant residual variance. This form of weighting is model-based and is called weighted least squares (WLS). Most of current statistical software packages, (e.g., SAS, Stata, S-Plus and R), use $(1 - R_{k(\text{WLS})}^2)^{-1}$ as VIF for WLS, where $R_{k(\text{WLS})}^2$ is the square of the multiple correlation from the WLS regression of the k^{th} column of \mathbf{X} on the other columns. Fox and Monette (1992) also generalized this concept of variance inflation as a measure of collinearity to a subset of parameters in \mathbf{b} and derived a *generalized variance-inflation factor* (GVIF). Furthermore, some interesting work has developed VIF-like measures, such as *collinearity indices* in Steward (1987) that are simply the square roots of the VIFs and *tolerance* defined as the inverse of VIF in Simon and Lesage (1988).

3. VIF in survey weighted least squares regression

3.1 Survey-weighted least squares estimators

Suppose the underlying structural model in the superpopulation is $\mathbf{Y} = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{e}$, where the error terms in the model have a general variance structure $\mathbf{e} \sim (0, \sigma^2 \mathbf{V})$ with known \mathbf{V} and σ^2 . Define \mathbf{W} to be the diagonal matrix of survey weights. We assume throughout that the survey weights are constructed in such a way that they can be used for estimating finite population totals. The survey weighted least squares (SWLS) estimator is $\hat{\boldsymbol{\beta}}_{\text{SW}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$, assuming $\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X}$ is invertible. Fuller (2002) describes the properties of this estimator.

The estimate $\hat{\boldsymbol{\beta}}_{\text{SW}}$ is a model unbiased estimator of $\boldsymbol{\beta}$ under the model $\mathbf{Y} = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{e}$ regardless of whether

$\text{Var}_M(\mathbf{e}) = \sigma^2 \mathbf{V}$ is specified correctly or not, and is approximately design-unbiased for the census parameter $\mathbf{B}_U = (\mathbf{X}_U^T \mathbf{X}_U)^{-1} \mathbf{X}_U^T \mathbf{Y}_U$, in the finite population of N units. The subscript U stands for the finite population, $\mathbf{Y}_U = (Y_1, \dots, Y_N)^T$, and $\mathbf{X}_U = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ with \mathbf{x}_k as the $N \times 1$ vector of values for covariate k .

3.2 Model variance of coefficient estimates

The model variance of the parameter estimator $\hat{\beta}_{\text{SW}}$, assuming $\text{Var}_M(\mathbf{e}) = \sigma^2 \mathbf{V}$, can be expressed as

$$\begin{aligned} \text{Var}_M(\hat{\beta}_{\text{SW}}) &= \sigma^2 (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{V}} \tilde{\mathbf{X}} (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \\ &= \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} \sigma^2 = \mathbf{G} \sigma^2, \end{aligned} \quad (2)$$

where $\tilde{\mathbf{X}} = \mathbf{W}^{1/2} \mathbf{X}$, $\tilde{\mathbf{V}} = \mathbf{W}^{1/2} \mathbf{V} \mathbf{W}^{1/2}$, $\mathbf{A} = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$, $\mathbf{B} = \tilde{\mathbf{X}}^T \tilde{\mathbf{V}} \tilde{\mathbf{X}}$, and $\mathbf{G} = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$.

If the columns of \mathbf{X} are orthogonal, then $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} = \text{diag}(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k)$ and $\mathbf{A}^{-1} = \text{diag}(1/\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k)$, where $\tilde{\mathbf{x}}_k = \mathbf{w}_k^{1/2} \mathbf{x}_k$. The ij^{th} element of \mathbf{G} then becomes $\tilde{\mathbf{x}}_i^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_j / (\tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_i)^2$. Thus, when the \mathbf{X} 's are orthogonal, the model variance of $\hat{\beta}_{\text{SW}_k}$ is

$$\text{Var}_M(\hat{\beta}_{\text{SW}_k}) = \sigma^2 \tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k / (\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k)^2, \quad (3)$$

a fact we will use later. More generally, the model variance of $\hat{\beta}_{\text{SW}_k}$, the coefficient estimate for the k^{th} explanatory variable, is

$$\text{Var}_M(\hat{\beta}_{\text{SW}_k}) = \mathbf{i}_k' \text{Var}_M(\hat{\beta}_{\text{SW}}) \mathbf{i}_k = \sigma^2 \mathbf{i}_k' \mathbf{G} \mathbf{i}_k = \sigma^2 g^{kk} \quad (4)$$

where \mathbf{i}_k is a $p \times 1$ vector with 1 in position k and 0's elsewhere, and g^{kk} is the k^{th} diagonal element of matrix \mathbf{G} .

3.3 Model-based VIF

As shown in Appendix A, the model variance of $\hat{\beta}_{\text{SW}_k}$ in (4) can be written as:

$$\text{Var}_M(\hat{\beta}_{\text{SW}_k}) = g^{kk} \sigma^2 = \frac{\zeta_k \rho_k}{1 - R_{\text{SW}(k)}^2} \frac{\sigma^2 \tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k}{(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k)^2}, \quad (5)$$

where

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{V}} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}} = \frac{\mathbf{e}_{xk}^T \mathbf{W} \mathbf{V} \mathbf{W} \mathbf{e}_{xk}}{\mathbf{e}_{xk}^T \mathbf{W} \mathbf{e}_{xk}},$$

with $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)} \hat{\beta}_{\text{SW}(k)}$ being the residual from SWLS regressing \mathbf{x}_k on $\mathbf{X}_{(k)}$ and $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)} = \mathbf{W}^{1/2} \mathbf{e}_{xk}$,

$$\rho_k = \frac{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k} = \frac{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W} \mathbf{V} \mathbf{W} \mathbf{x}_k}$$

and $R_{\text{SW}(k)}^2$, defined in Appendix A, is the square of the multiple correlation from the weighted regression of the k^{th} column of \mathbf{X} on the other columns. Hence, ζ_k and ρ_k

depend on \mathbf{W} and \mathbf{V} . The variance under orthogonality in (3) is inflated

$$\text{VIF}_k = \frac{\zeta_k \rho_k}{1 - R_{\text{SW}(k)}^2} \quad (6)$$

times when incorporating the other $p-1$ explanatory variables in SWLS. The model-based VIF in SWLS includes not only the multiple correlation coefficient $R_{\text{SW}(k)}^2$ but also two adjustment coefficients, ζ_k and ρ_k , that are not present in the OLS and WLS cases.

Using the singular value decomposition of $\tilde{\mathbf{V}}$, we can bound the factor $\zeta_k \rho_k$, which is the adjustment to the VIF in WLS. Based on the extrema of the ratio of quadratic forms (Lin 1984), the term ζ_k is bounded in the range of $\mu_{\min}(\tilde{\mathbf{V}}) \leq \zeta_k \leq \mu_{\max}(\tilde{\mathbf{V}})$, and ρ_k is bounded in the range of

$$\frac{1}{\mu_{\max}(\tilde{\mathbf{V}})} \leq \rho_k \leq \frac{1}{\mu_{\min}(\tilde{\mathbf{V}})},$$

where $\mu_{\min}(\tilde{\mathbf{V}})$ and $\mu_{\max}(\tilde{\mathbf{V}})$ are the minimum and maximum singular values of the matrix $\tilde{\mathbf{V}}$. Combining these results, the joint coefficient $\zeta_k \rho_k$ is bounded in the range of:

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \leq \zeta_k \rho_k \leq \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

Notice that when $\tilde{\mathbf{V}} = \mathbf{I}$, $\zeta_k = \rho_k = 1$ and (6) reduces to

$$\frac{1}{1 - R_{\text{SW}(k)}^2} \frac{\sigma^2}{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k},$$

which is the model variance of the WLS estimates when \mathbf{V} is diagonal and \mathbf{W} is correctly specified as $\mathbf{W} = \mathbf{V}^{-1}$. In that unusual case, the VIF currently computed by software packages will be appropriate for SWLS. However, rarely will it be reasonable to think that $\mathbf{W} = \mathbf{V}^{-1}$ in survey estimation. If $\tilde{\mathbf{V}} \neq \mathbf{I}$, then ζ_k and ρ_k are not equal to 1 and a specialized calculation of the VIF is still needed. When $\mathbf{V} = \mathbf{I}$, which is the usual application considered by analysts,

$$\tilde{\mathbf{V}} = \mathbf{W}, \zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T \mathbf{W} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}}, \rho_k = \frac{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T \mathbf{W} \tilde{\mathbf{x}}_k}$$

and

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}},$$

where w_{\min} is the minimum value of the survey weights and w_{\max} is their maximum value. In this case, the range of $\zeta_k \rho_k$ is bounded by

$$\left[\frac{w_{\min}}{w_{\max}}, \frac{w_{\max}}{w_{\min}} \right].$$

When all the survey weights are constant, $\zeta_k \rho_k = 1$ and the VIF produced by standard software, $(1 - R_{SW(k)}^2)^{-1}$, does not need to be adjusted in SWLS; however, when the range of the survey weights is large, $\zeta_k \rho_k$ can be very small or large and can be either above or below 1. In this case the VIF produced by standard software is not appropriate and a special calculation is needed. These facts will be illustrated in our experimental studies.

The VIF in (6) is appropriate regardless of whether the model contains an intercept or not. An alternative version can also be written that assumes that an intercept is in the model when $\tilde{\mathbf{x}}_k$ is regressed on the other \mathbf{x} 's. The derivation of this form is in Liao (2010). We summarize the result below.

The variance of $\hat{\beta}_{SW_k}$ in a model M2 that includes an intercept and in which $\tilde{\mathbf{x}}_k$ is orthogonal to the other \mathbf{x} 's is:

$$\text{Var}_{M2}(\hat{\beta}_{SW_k}) = \frac{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{I}}\tilde{\bar{x}}_k)^T \tilde{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{I}}\tilde{\bar{x}}_k)}{\text{SST}_{SW_m(k)}^2} \quad (7)$$

where $\tilde{\mathbf{I}} = (w_1^{1/2}, \dots, w_n^{1/2})$, $\tilde{\bar{x}}_k = \sum_{i \in s} w_i x_{ki} / \hat{N}$, $\hat{N} = \sum_{i \in s} w_i$, and $\text{SST}_{SW_m(k)} = \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N} \tilde{\bar{x}}_k^2$. The variance of $\hat{\beta}_{SW_k}$ can then be rewritten as

$$\text{Var}_M(\hat{\beta}_{SW_k}) = \frac{\zeta_k \rho_{mk}}{1 - R_{SW_m(k)}^2} \text{Var}_{M2}(\hat{\beta}_{SW_k}) \quad (8)$$

where $R_{SW_m(k)}^2$ is the SWLS R-square from regressing $\tilde{\mathbf{x}}_k$ on the \mathbf{x} 's in the remainder of $\tilde{\mathbf{X}}$ (excluding a column for the intercept). The term ζ_k was defined following (5) and

$$\rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N} \tilde{\bar{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{I}}\tilde{\bar{x}}_k)^T \tilde{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{I}}\tilde{\bar{x}}_k)}.$$

Most software packages will consistently provide $(1 - R_{SW_m(k)}^2)^{-1}$ as the VIF as part of WLS regression output. Note that this is different from the VIF, $(1 - R_{SW(k)}^2)^{-1}$, introduced in section 3.3 which does not assume that an intercept is retained in the model. Software packages generally do not supply $(1 - R_{SW(k)}^2)^{-1}$.

Using arguments similar to those in the previous section, we can bound $\zeta_k \rho_{mk}$ by

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \leq \zeta_k \rho_{mk} \leq \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

The model variance of $\hat{\beta}_{SW_k}$ is inflated by

$$\text{VIF}_{mk} = \frac{\zeta_k \rho_{mk}}{1 - R_{SW_m(k)}^2}$$

compared to its variance in the model (M2) with only the explanatory variable $\tilde{\mathbf{x}}_k$ and intercept. The new intercept-adjusted VIF_{mk} retains some properties of VIF_k in (6).

When $\tilde{\mathbf{V}} = \mathbf{I}$, we have $\zeta_k = 1$, $\rho_{mk} = 1$ and the intercept-adjusted VIF in (8) for SWLS is equal to the conventional intercept-adjusted VIF: $(1 - R_{m(k)}^2)^{-1}$. When $\mathbf{V} = \mathbf{I}$, we have $\tilde{\mathbf{V}} = \mathbf{W}$,

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T \mathbf{W} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}}, \rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N} \tilde{\bar{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{I}}\tilde{\bar{x}}_k)^T \mathbf{W}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{I}}\tilde{\bar{x}}_k)}$$

and

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}}.$$

The range of $\zeta_k \rho_{mk}$ also depends on the range of survey weights as did $\zeta_k \rho_k$.

3.4 Estimating the VIF for a model with stratified clustering when \mathbf{V} is unknown

In the previous sections, we used model-based arguments to derive VIFs. The VIFs contain terms, $\tilde{\mathbf{V}}$ in particular, that are unknown and must be estimated. In this section, we construct estimators of the components of the VIFs, again using model-based arguments. However, a standard, design-based linearization variance estimator also estimates the model variance, as shown below, and supplies the components needed to estimate the VIF. In the remainder of this section, we will present estimators that are appropriate for a model that has a stratified clustered covariance structure.

Suppose that in a stratified multistage sampling design, there are $h = 1, \dots, H$ strata in the population, $i = 1, \dots, N_h$ clusters in the corresponding stratum h and $t = 1, \dots, M_{hi}$ units in cluster hi . We select $i = 1, \dots, n_h$ clusters in stratum h and $t = 1, \dots, m_{hi}$ units in cluster hi . Denote the set of sample clusters in stratum h by s_h and the sample of units in cluster hi as s_{hi} . The total number of sample units in stratum h is $m_h = \sum_{i \in s_h} m_{hi}$, and the total in the sample is $m = \sum_{h=1}^H m_h$. Clusters are assumed to be selected with replacement within strata and independently between strata. Consider this model:

$$\begin{aligned} E_M(Y_{hit}) &= \mathbf{x}_{hit}^T \boldsymbol{\beta} \\ h &= 1, \dots, H, \quad i = 1, \dots, N_h, \quad t = 1, \dots, M_{hi} \\ \text{Cov}_M(Y_{hit}, Y_{h't'}) &= 0 \\ h &\neq h', \text{ or, } h = h' \text{ and } i \neq i'. \end{aligned} \quad (9)$$

Units within each cluster are assumed to be correlated but the particular correlation of the covariances does not have to be specified for this analysis. The estimator of the regression parameter is:

$$\hat{\boldsymbol{\beta}}_{SW} = \sum_{h=1}^H \sum_{i \in s_h} \mathbf{A}^{-1} \mathbf{x}_{hi}^T \mathbf{W}_{hi} \mathbf{Y}_{hi} \quad (10)$$

where \mathbf{X}_{hi} is the $m_{hi} \times p$ matrix of covariates for sample units in cluster hi , $\mathbf{W}_{hi} = \text{diag}(w_t)$, $t \in s_{hi}$ is the diagonal

matrix of survey weights for cluster hi and \mathbf{Y}_{hi} is the $m_{hi} \times 1$ vector of response variables in cluster hi . The model variance of $\hat{\beta}_{sw}$ is:

$$\begin{aligned} \text{Var}_M(\hat{\beta}_{sw}) &= \mathbf{A}^{-1} \left[\sum_{h=1}^H \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{V}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi} \right] \mathbf{A}^{-1} \\ &= \mathbf{A}^{-1} \left[\sum_{h=1}^H \mathbf{X}_h^T \mathbf{W}_h \mathbf{V}_h \mathbf{W}_h \mathbf{X}_h \right] \mathbf{A}^{-1}, \quad (11) \end{aligned}$$

where $\mathbf{V}_{hi} = \text{Var}_M(\mathbf{Y}_{hi})$ and $\mathbf{V}_h = \text{Blkdiag}(\mathbf{V}_{hi})$, $i \in s_h$. Expression (11) is a special case of (2) with $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T, \dots, \mathbf{X}_H^T)$, \mathbf{X}_h is the $m_h \times p$ matrix of covariates for sample units in stratum h , $\mathbf{W} = \text{diag}(\mathbf{W}_{hi})$, for $h = 1, \dots, H$ and $i \in s_h$ and $\mathbf{V} = \text{Blkdiag}(\mathbf{V}_h)$.

Denote the cluster-level residuals as a vector, $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\beta}_{sw}$. A design-based linearization estimator is:

$$\begin{aligned} \text{var}_L(\hat{\beta}_{sw}) &= \mathbf{A}^{-1} \left[\sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{i \in s_h} (\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)(\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)^T \right] \mathbf{A}^{-1} \\ &= \mathbf{A}^{-1} \left[\sum_{h=1}^H \frac{n_h}{n_h - 1} \left(\sum_{i \in s_h} \mathbf{z}_{hi} \mathbf{z}_{hi}^T - n_h \bar{\mathbf{z}}_h \bar{\mathbf{z}}_h^T \right) \right] \mathbf{A}^{-1}, \quad (12) \end{aligned}$$

where

$$\bar{\mathbf{z}}_h = \frac{1}{n_h} \sum_{i \in s_h} \mathbf{z}_{hi}$$

and $\mathbf{z}_{hi} = \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{e}_{hi}$ with $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\beta}_{sw}$. This expression can be reduced to the formula for a single-stage stratified design when the cluster sample sizes are all equal to 1, $m_{hi} = 1$. Expression (12) is used by the Stata and SUDAAN packages, among others. The estimator $\text{var}_L(\hat{\beta}_{sw})$ is consistent and approximately design-unbiased under a design where clusters are selected with replacement (Fuller 2002). Li (2007a,b) showed that (12) is also an approximately model-unbiased estimator under model (11).

The term in brackets in (12) serves as an estimator of the matrix \mathbf{B} in expression (2). The components of $\text{var}_L(\hat{\beta}_{sw})$ can be used to construct estimators of ζ_k and ρ_k in (5) and ρ_{mk} in (8). In particular,

$$\hat{\zeta}_k = \frac{\mathbf{e}_{xk}^T \hat{\mathbf{W}} \hat{\mathbf{V}} \mathbf{W} \mathbf{e}_{xk}}{\mathbf{e}_{xk}^T \mathbf{W} \mathbf{e}_{xk}}, \quad (13)$$

where

$$\hat{\mathbf{V}} = \text{Blkdiag} \left[\frac{n_h}{n_h - 1} \left(\hat{\mathbf{V}}_h - \frac{1}{n_h} \mathbf{e}_h \mathbf{e}_h^T \right) \right], \quad h = 1, 2, \dots, H,$$

with $\hat{\mathbf{V}}_h = \text{Blkdiag}(\mathbf{e}_{hi} \mathbf{e}_{hi}^T)$ and

$$\hat{\rho}_k = \frac{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k}{\mathbf{x}_k^T \hat{\mathbf{W}} \mathbf{W} \mathbf{x}_k},$$

with $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)} \hat{\beta}_{sw(k)}$. The estimate of $\hat{\rho}_{mk}$, defined following (8), is

$$\hat{\rho}_{mk} = \frac{(\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k - \hat{N} \bar{\mathbf{x}}_k^2)}{(\mathbf{x}_k - \mathbf{1} \bar{\mathbf{x}}_k)^T \hat{\mathbf{W}} \mathbf{V} \mathbf{W} (\mathbf{x}_k - \mathbf{1} \bar{\mathbf{x}}_k)}. \quad (14)$$

Given these component estimators VIF_k is estimated by

$$\widehat{\text{VIF}}_k = \frac{\hat{\zeta}_k \hat{\rho}_k}{1 - R_{sw(k)}^2}$$

and VIF_{mk} is estimated by

$$\widehat{\text{VIF}}_{mk} = \frac{\hat{\zeta}_k \hat{\rho}_{mk}}{1 - R_{swm(k)}^2}.$$

4. Experimental study

We will now illustrate the proposed, modified collinearity diagnostics and investigate their behavior using dietary intake data from the National Health and Nutrition Examination Survey (NHANES) 2007-2008 (http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/datadoc_changes_0708.htm). The dietary intake data are used to estimate the types and amounts of foods and beverages consumed during the 24-hour period prior to the interview (midnight to midnight), and to estimate intakes of energy, nutrients, and other food components from those foods and beverages. NHANES uses a complex, multistage, probability sampling design. Oversampling of certain population subgroups is done to increase the reliability and precision of health status indicator estimates for these groups. Among the respondents who received the in-person interview in the mobile examination center (MEC), around 94% provided complete dietary intakes. The survey weights in this data were constructed by taking MEC sample weights and further adjusting for the additional nonresponse and the differential allocation by day of the week for the dietary intake data collection. These weights are more variable than the MEC weights. The data set used in our study is a subset of 2007-2008 data composed of female respondents aged 26 to 40. Observations with missing values in the selected variables are excluded from the sample which finally contains 672 complete respondents. The final weights in our sample range from 6,028 to 330,067, with a ratio of 55:1. The U.S. National Center for Health Statistics recommends that the design of the sample is approximated by the stratified selection with replacement of 32 PSUs from 16 strata, with 2 PSUs within each stratum.

For this empirical study, a linear regression of body weight(kg) is fitted using survey weighted least squares. The predictor variables considered include age, Black(race) and

nine daily total nutrition intake variables, which are calorie(100kcal), protein(100gm), carbohydrate(100gm), sugar(100gm), total fat(100gm), total saturated fatty acids(100gm), total monounsaturated fatty acids(100gm), total polyunsaturated fatty acids(100gm) and alcohol(100gm). All the daily total nutrition intake variables are correlated with each other to different degrees as shown in Figure 1.

Three regression methods compared in this study. The first one uses *ordinary least squares* (OLS) method and ignores sampling complexities including the weighting. The second one uses *weighted least squares* (WLS), which incorporates the survey weights by assuming $\mathbf{V} = \mathbf{W}^{-1}$ but ignores all sampling complexities. The third one is *survey weighted least squares* (SWLS), which uses the actual complex sampling design as described in section 3.4. The weight matrices, coefficient variance estimators and collinearity diagnostics of these three methods are listed in Table 1.

The results from fitting the model using three different regression methods are displayed in Table 2. The model with all the predictors is shown in the upper part of the table. In the lower tier of the table, a reduced model with less of the near-dependency problem is fitted with only three predictors: age, Black and calorie. In the reduced model, the value of the coefficient for calorie is positive and significant when WLS or SWLS is used, which seems logical and reflects the anticipated positive relationship between a respondent's body weight and her daily total calorie intake. However, when the other total nutrition intake variables are included in the model, the value of the calorie coefficient is negative and not significant due to its inflated variance. This is a typical example in which the variance of a coefficient is inflated, and its sign is illogical due to collinearity.

Table 3 reports the VIF values when the three different regression methods are used. The VIF formulas for these regression methods are listed in Table 1. When all the predictors are included in the model, calorie has the largest VIF values in all the regressions due to its high near-dependency with all the other total nutrition intake variables. As shown in Table 1, the VIF in SWLS can be obtained by multiplying the VIF from WLS with the adjustment coefficient $\zeta_k \hat{\rho}_k$. In Table 3, the adjustment coefficients $\zeta_k \hat{\rho}_k$ for all the non-fat total nutrition intake variables are all less than 1, especially the one for carbohydrate which is 0.46. This indicates that the VIF values for these variables in SWLS are much smaller than the ones in WLS and the collinearity among predictors in the model has less impact on the coefficient estimation when using SWLS, compared to using WLS. But for the fat-related nutrition intake variables, their $\zeta_k \hat{\rho}_k$ are all larger than 1. Thus, the collinearity among the fat-related nutrition intake variables is more harmful to the coefficient estimation in SWLS than in WLS. To take a closer look at this problem, we also fitted a model that only contains two nutrition intake variables: total fat and total monounsaturated fatty acids. The SWLS VIF values are three times as large as the ones from OLS or WLS for these two nutrition variables. If an analyst is analyzing this survey data using SWLS but uses the unadjusted VIF values provided by standard statistical packages for either OLS (as shown in the first column) or WLS (as shown in the second column), the unadjusted VIFs will give somewhat misleading judgements on the severity of collinearity in this model. In summary, although the estimated slopes and predictions in regression using WLS and SWLS are the same, the VIFs can be underestimated or overestimated if survey complexities are ignored.

Table 1
Regression methods and their collinearity diagnostic statistics used in this experimental study

Regression Type	Weight Matrix \mathbf{W}^a	Variance Estimation of $\hat{\beta}$	VIF fomula
OLS	\mathbf{I}	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$	$\text{VIF} = \frac{1}{1 - R_{m(k)}^2}$
WLS	\mathbf{W}^c	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\text{VIF} = \frac{1}{1 - R_{SWm(k)}^2}$
SWLS	\mathbf{W}	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \hat{\mathbf{V}} \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$	$\text{VIF} = \frac{\hat{\zeta}_k \hat{\rho}_{mk}}{1 - R_{SWm(k)}^2}$
		with	$\text{with } \hat{\zeta}_k = \frac{\mathbf{e}_{xk}^T \mathbf{W} \hat{\mathbf{V}} \mathbf{W} \mathbf{e}_{xk}}{\mathbf{e}_{xk}^T \mathbf{W} \mathbf{e}_{xk}},$
		$\hat{\mathbf{V}} = \sum_{h=1}^H \frac{n_h}{n_h - 1} \left[\text{Blkdiag}(\mathbf{e}_{hi} \mathbf{e}_{hi}^T) - \frac{1}{n_h} \mathbf{e}_h \mathbf{e}_h^T \right]$	$\hat{\rho}_{mk} = \frac{(\bar{\mathbf{x}}_k^T \bar{\mathbf{x}}_k - \bar{N} \bar{\mathbf{x}}_k^2)}{(\bar{\mathbf{x}}_k - \bar{\mathbf{1}} \bar{\mathbf{x}}_k)^T \hat{\mathbf{V}} (\bar{\mathbf{x}}_k - \bar{\mathbf{1}} \bar{\mathbf{x}}_k)}$

^a In all the regression models, the parameters are estimated by: $\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$.

^b $R_{m(k)}^2$ is the OLS R-square from regressing \mathbf{x}_k on the \mathbf{x} 's in the remainder of \mathbf{X} (excluding a column for the intercept).

^c \mathbf{W} is the diagonal matrix with survey weights w_i on the main diagonal.



^a T.Fat: total fat;
T.S.Fat: total saturated fatty acid;
T.M.Fat: total monounsaturated fatty acid;
T.P.Fat: total polyunsaturated fatty acid.

Figure 1 Pairwise scatterplots and correlation coefficients of nutrition variables^a

Table 2
Parameter estimates with their associated standard errors using three different regression methods

Variable	OLS		Full Model WLS		SWLS	
	Beta	SE.	Beta	SE.	Beta	SE.
Intercept	63.90*** ^a	6.95	67.47***	6.36	67.47***	8.76
Age	0.26	0.19	0.08	0.18	0.08	0.25
Black	10.39***	2.07	10.59***	2.38	10.59***	2.20
Calorie	-6.41	5.76	-8.19	5.56	-8.19	5.75
Protein	25.72	24.76	40.98	23.60	40.98	25.38
Carbohydrate	26.67	23.93	32.31	22.96	32.31	22.65
Sugar	-1.90	3.06	-0.30	2.82	-0.30	4.06
Fiber	-41.17	20.23	-34.20	17.98	-34.20	19.05
Alcohol	38.84	39.45	49.37	38.28	49.37	40.10
Total Fat	150.25*	69.53	161.78*	72.12	161.78	94.76
Total Saturated Fatty Acids	-113.20*	49.81	-101.40	56.26	-101.40	82.71
Total Monounsaturated Fatty Acids	-72.05	48.03	-92.44	51.52	-92.44	83.51
Total Polyunsaturated Fatty Acids	-92.60*	46.13	-75.55	51.16	-75.55	78.76

Variable	OLS		Reduced Model WLS		SWLS	
	Beta	SE.	Beta	SE.	Beta	SE.
Intercept	62.26***	6.88	67.52***	6.29	67.52***	8.48
Age	0.27	0.19	0.07	0.18	0.07	0.25
Black	12.54***	1.98	11.74***	2.32	11.74***	2.05
Calorie	0.15	0.10	0.23*	0.09	0.23*	0.10

^a p values of significance: * p = 0.05; ** p = 0.01; *** p = 0.005.

Table 3
VIF values using three different regression methods

Variable	Full Model			
	OLS VIF	WLS VIF	SWLS VIF	$\zeta_k \rho_k$
Age	1.02	1.03	0.96	0.94
Black	1.10	1.07	1.12	1.05
Calorie	3,411.61	3,562.70	2,740.83	0.77
Protein	123.12	127.35	103.50	0.81
Carbohydrate	1,074.87	1,007.40	462.08	0.46
Sugar	8.37	7.03	4.87	0.69
Fiber	4.59	3.94	2.37	0.60
Alcohol	120.56	115.67	89.92	0.78
Total Fat	1,190.24	1,475.27	2,513.69	1.70
Total Saturated Fatty Acids	76.80	112.61	202.91	1.80
Total Monounsaturated Fatty Acids	82.37	107.34	286.24	2.67
Total Polyunsaturated Fatty Acids	34.73	49.45	118.21	2.39

Variable	Reduced Model			
	OLS VIF	WLS VIF	SWLS VIF	$\zeta_k \rho_k$
Age	1.00	1.00	0.98	0.98
Black	1.02	1.01	0.97	0.96
Total Fat	20.10	20.22	63.15	3.12
Total Monounsaturated Fatty Acids	20.16	20.26	61.57	3.04

Variable	Reduced Model			
	OLS VIF	WLS VIF	SWLS VIF	$\zeta_k \rho_k$
Age	1.00	1.00	0.98	0.97
Black	1.00	1.03	1.00	1.00
Calorie	1.00	1.01	0.96	0.95

5. Conclusion

Regression diagnostics need to be adapted to be appropriate for models estimated from survey data to account for the use of weights and design features like stratification and clustering. In this paper we developed a new formulation for a variance inflation factor (VIF) appropriate for linear models. A VIF measures the amount by which the variance of a parameter estimator is inflated due to predictor variables being correlated with each other, rather than being orthogonal. Although survey-weighted regression slope estimates can be obtained from weighted least squares procedures in standard software packages, the VIFs produced by the non-survey routines are incorrect. The complex sample VIF is equal to the VIF from weighted least squares times an adjustment factor. The adjustment factor is positive but can be either larger or smaller than 1, depending on the nature of the data being analyzed.

In an empirical study, we illustrated the application of our new approach using data from the 2007-2008 National Health and Nutrition Examination Survey. We provided a simple example of how the collinearity among predictors affects the estimation of coefficients in linear regression and

demonstrated that although the estimated coefficients (and fitted values) are the same when weighted least squares or survey-weighted least squares are used, their estimated variances and VIF values (reflecting the impact of collinearity on coefficient estimation) can be different.

The goals of an analysis must be considered in deciding how to use VIFs. If prediction is the main objective, then including collinear variables or selecting incorrect variables is less of a concern. If more substantive conclusions are desired, then an analyst should consider which variables should logically be included as predictors rather than relying on some automatic algorithm for variable selection. VIFs are a useful tool for identifying predictors whose estimated coefficients have variances that are unnecessarily large. Although VIFs might be considered as a tool for automatic variable selection, simulations in Liao (2010), not reported here, show that using VIFs is not a reliable way of identifying a true underlying model.

Acknowledgements

The authors thank the associate editor and referees whose comments led to important improvements.

Appendix A

Derivation of g^{kk}

Similar to the derivation of conventional OLS VIF in Theil (1971), the sum of squares and cross products matrix $\mathbf{A} = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$, which can be partitioned as

$$\mathbf{A}_{p \times p} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T \tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{X}}_{(k)} \end{pmatrix} \quad (15)$$

where the columns of $\tilde{\mathbf{X}}$ are reordered so that $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_k \tilde{\mathbf{X}}_{(k)})$ with $\tilde{\mathbf{X}}_{(k)}$ being the $n \times (p-1)$ matrix containing all columns except the k^{th} column of $\tilde{\mathbf{X}}$.

Using the formula for the inverse of a partitioned matrix, the upper-left element of \mathbf{A}^{-1} can be expressed as:

$$\begin{aligned} a^{kk} &= \mathbf{i}_k^T \mathbf{A}^{-1} \mathbf{i}_k = \mathbf{i}_k^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{i}_k \\ &= \frac{1}{(1 - R_{\text{SW}(k)}^2) \text{SST}_{\text{SW}(k)}} \\ &= \frac{1}{(1 - R_{\text{SW}(k)}^2) \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \end{aligned} \quad (16)$$

where

$$R_{\text{SW}(k)}^2 = \frac{\hat{\beta}_{\text{SW}(k)}^T \tilde{\mathbf{x}}_{(k)}^T \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)}}{\text{SST}_{\text{SW}(k)}},$$

with $\hat{\beta}_{\text{SW}(k)} = (\tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{X}}_{(k)})^{-1} \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{x}}_k$, is the coefficient of determination corresponding to the regression of $\tilde{\mathbf{x}}_k$ on the $p-1$ other explanatory variables. The term $\text{SST}_{\text{SW}(k)} = \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k$, is the total sum of squares in this regression.

The term $(1 - R_{\text{SW}(k)}^2)^{-1}$ in (16) is the VIF that will be produced by standard statistical packages when a weighted least squares regression is run. Under the model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $\epsilon \sim (0, \sigma^2 \mathbf{W}^{-1})$, expression (16) is equal to $\text{Var}_M(\hat{\beta}_{\text{SW}(k)})/\sigma^2$. However, this is not appropriate for survey-weighted least squares regressions because the variance of $\hat{\beta}_{\text{SW}}$ has the more complex form in (2).

The matrix $\mathbf{G} = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$ can be expressed as:

$$\mathbf{G} = \begin{pmatrix} a^{kk} & \mathbf{a}^{(k)(k)} \\ \mathbf{a}^{(k)(k)} & \mathbf{A}^{(k)(k)} \end{pmatrix} \begin{pmatrix} b_{kk} & \mathbf{b}_{(k)(k)} \\ \mathbf{b}_{(k)(k)} & \mathbf{B}_{(k)(k)} \end{pmatrix} \begin{pmatrix} a^{kk} & \mathbf{a}^{(k)(k)} \\ \mathbf{a}^{(k)(k)} & \mathbf{A}^{(k)(k)} \end{pmatrix} \quad (17)$$

where the inverse matrix is $\mathbf{A}^{-1} = [a^{hk}]$, $h, k = 1, \dots, p$, $\mathbf{a}^{(k)(k)}$ is defined as the k^{th} row of \mathbf{A}^{-1} excluding a^{kk} , $(a^{k1}, \dots, a^{k(k-1)}, a^{k(k+1)}, \dots, a^{kp})$, $\mathbf{a}^{(k)(k)} = [\mathbf{a}^{(k)(k)}]^T$ and $\mathbf{A}^{(k)(k)}$ is defined as the $(k-1) \times (k-1)$ part of matrix \mathbf{A}^{-1} excluding the k^{th} row and column. The partitioned version of \mathbf{B} is

$$\mathbf{B} = \begin{pmatrix} b_{kk} & \mathbf{b}_{(k)(k)} \\ \mathbf{b}_{(k)(k)} & \mathbf{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \end{pmatrix}. \quad (18)$$

By virtue of the symmetry of \mathbf{A} and \mathbf{B} , the k^{th} diagonal element of \mathbf{G} is

$$g^{kk} = a^{kk} (a^{kk} b_{kk} + 2\mathbf{b}_{(k)(k)} \mathbf{a}^{(k)(k)} + \mathbf{a}^{(k)(k)T} \mathbf{B}_{(k)(k)} \mathbf{a}^{(k)(k)}). \quad (19)$$

Using the partitioned inverse of matrix \mathbf{A} , which represents $(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}$, it can be shown that

$$\mathbf{a}^{(k)(k)} = -a^{kk} (\tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{x}}_k)^{-1} \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k = -a^{kk} \hat{\beta}_{\text{SW}(k)}. \quad (20)$$

Substituting $a^{(k)(k)}$ in (19), g^{kk} can be compactly expressed in terms of a^{kk} , $\hat{\beta}_{\text{SW}(k)}$ and the lower right component of matrix \mathbf{B} :

$$\begin{aligned} g^{kk} &= (a^{kk})^2 (b_{kk} - 2\mathbf{b}_{(k)(k)} \hat{\beta}_{\text{SW}(k)} + \hat{\beta}_{\text{SW}(k)}^T \mathbf{B}_{(k)(k)} \hat{\beta}_{\text{SW}(k)}) \\ &= a^{kk} \times \frac{1}{1 - R_{\text{SW}(k)}^2} \frac{1}{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \\ &\quad \times (\tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k - 2\tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)} + \hat{\beta}_{\text{SW}(k)}^T \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)}) \\ &= a^{kk} \times \frac{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)})^T \tilde{\mathbf{V}} (\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)})}{(1 - R_{\text{SW}(k)}^2) \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \\ &= a^{kk} \times \frac{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)})^T \tilde{\mathbf{V}} (\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)})}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)})^T (\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)})} \\ &= \frac{1}{1 - R_{\text{SW}(k)}^2} \frac{1}{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \frac{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{V}} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}}, \end{aligned} \quad (21)$$

where $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{SW}(k)}$ is the residual from regressing $\tilde{\mathbf{x}}_k$ on $\tilde{\mathbf{X}}_{(k)}$.

References

- Belsley, D.A. (1984). Collinearity and forecasting. *Journal of Forecasting*, 38, 73-93.
- Belsley, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York: John Wiley & Sons, Inc.
- Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York: Wiley Interscience.
- Elliot, M. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 23-34.
- Farrar, D.E., and Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.
- Fox, J. (1984). *Linear Statistical Models and Related Methods, With Applications to Social Research*. New York: John Wiley & Sons, Inc.
- Fox, J., and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417), 178-183.

- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28(1), 5-23.
- Kmenta, J. (1986). *Elements of Econometrics*. New York: Macmillan, 2nd Ed.
- Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3341-3348.
- Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland. Available at <http://drum.lib.umd.edu/bitstream/1903/7598/1/umi-umd-4863.pdf>.
- Li, J., and Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, 35(1), 15-24.
- Li, J., and Valliant, R. (2011). Linear regression influence diagnostics for unclustered survey data. *Journal of Official Statistics*, 20, 99-119.
- Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Unpublished doctoral dissertation, University of Maryland. Available at http://drum.lib.umd.edu/bitstream/1903/10881/1/Liao_umd_0117E_11537.pdf.
- Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.
- Neter, J., Kutner, M., Wasserman, W. and Nachtsheim, C. (1996). *Applied Linear Statistical Models*. New York: McGraw-Hill/Irwin, 4th Ed.
- Simon, S.D., and Lesage, J.P. (1988). The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management*, 1, 137-152.
- Smith, G. (1974). Multicollinearity and forecasting. Yale University Cowles Foundation Discussion Paper No.383. Available at <http://cowles.econ.yale.edu/P/cd/d03b/d0383.pdf>.
- Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.
- Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons, Inc.

Estimating agreement coefficients from sample survey data

Hung-Mo Lin, Hae-Young Kim, John M. Williamson and Virginia M. Lesser¹

Abstract

We present a generalized estimating equations approach for estimating the concordance correlation coefficient and the kappa coefficient from sample survey data. The estimates and their accompanying standard error need to correctly account for the sampling design. Weighted measures of the concordance correlation coefficient and the kappa coefficient, along with the variance of these measures accounting for the sampling design, are presented. We use the Taylor series linearization method and the jackknife procedure for estimating the standard errors of the resulting parameter estimates. Body measurement and oral health data from the Third National Health and Nutrition Examination Survey are used to illustrate this methodology.

Key Words: Clustering; Concordance correlation coefficient; Generalized estimating equations; Jackknife estimator; Kappa coefficient; Sample weighting; Stratification; Taylor series linearization.

1. Introduction

Surveys often collect multiple measures of latent conditions such as quality of life and aspiration for a college education, as well as multiple measures of difficult-to-classify conditions such as having chronic fatigue syndrome. When multiple measures are collected, interest naturally focuses on the agreement between the multiple measures and in obtaining confidence intervals on those agreement measures. Also, there may be interest in contrasting agreement across population subgroups and across alternate pairings of measurements. In this context, one might be interested in testing equality of agreement measures. This paper focuses on two measures of agreement between such multiple measures, the concordance correlation coefficient (CCC, ρ_c) and the kappa (κ) coefficient. The former is useful for continuous measurements with natural scales. If a measure of a latent concept has no natural scale, then it can be arbitrarily rescaled to have mean zero and unit variance. When this is possible, it is meaningless to talk about differences in marginal moments. However, if there is a natural scale, then rescaling is not desirable and a good measure of agreement will take into account both correlation and agreement of marginal moments. The kappa coefficient is most useful for binary classifications.

The CCC has been shown to be more appropriate for measuring agreement or reproducibility (Lin 1989; Lin 1992) than the Pearson correlation coefficient (ρ). It evaluates the accuracy between two readings by measuring the variation of the fitted linear relationship from the 45° line through the origin (the concordance line) and precision by measuring how far each observation deviates from the fitted

line. Let Y_{i1} and Y_{i2} denote a pair of continuous random variables measured on the same subject i using two methods. The CCC for measuring the agreement of Y_{i1} and Y_{i2} is defined as follows:

$$\rho_c = 1 - \frac{E[(Y_{i1} - Y_{i2})^2]}{E_{\text{indep}}[(Y_{i1} - Y_{i2})^2]} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \quad (1)$$

where $\sigma_1^2 = \text{var}(Y_{i1})$, $\sigma_2^2 = \text{var}(Y_{i2})$, and $\sigma_{12} = \text{cov}(Y_{i1}, Y_{i2})$ (Lin 1989). As noted by Lin (1989), $\rho_c = 0$ if and only if $\rho = 0$. It can also be shown algebraically that ρ_c is proportional to ρ and that $-1 \leq -|\rho| \leq \rho_c \leq |\rho| \leq 1$ (Lin 1989). Hence imprecision can be reflected by a smaller ρ and systematic bias can be reflected by a smaller ratio of ρ_c relative to ρ . Together, information on ρ and ρ_c provide a set of tools to identify which corrective actions, either to improve accuracy and/or to improve precision, is most beneficial (Lin and Chinchilli 1997).

The intraclass correlation coefficient (ICC) is also a popular measure of agreement for variables measured on a continuous scale (Fleiss 1986). Suppose Y_{i1} and Y_{i2} can be described in a linear model as follows: $y_{ij} = \mu_j + \theta_i + e_{ij}$ where μ_j is the mean of the measurement from the j^{th} method, $\theta_i \sim (0, \sigma_\theta^2)$ is the latent variable for the i^{th} subject, and the $e_{ij} \sim (0, \sigma_e^2)$ are independent errors terms. Carrasco and Jover (2003, page 850) used a model with variance components to demonstrate that the CCC is the intraclass correlation coefficient (ICC) when one takes into account the difference in averages of the methods:

$$\rho_{\text{ICC}} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_e^2 + \sigma_\mu^2} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}.$$

1. Hung-Mo Lin, Department of Anesthesiology, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1010, New York, NY 10029, U.S.A. E-mail: hung-mo.lin@mssm.edu; Hae-Young Kim, Center for Statistical Analysis and Research, New England Research Institutes, 9 Galen Street, Watertown, MA 02472, U.S.A.; John M. Williamson, Center for Global Health Research, Centers for Disease Control and Prevention/Kenya Medical Research Institute, 1578 Kisumu-Busia Road, Kisian Kisumu, Kenya; Virginia M. Lesser, Department of Statistics and Survey Research Center, Oregon State University, 44 Kidder Hall, Corvallis, OR 97331-4606, U.S.A.

Therefore, one can estimate the CCC using the variance components of a mixed effects model or the common method of moments. Because of its superiority to the Pearson correlation coefficient and its link to the ICC, application of the CCC has gained popularity in recent years (Chinchilli, Martel, Kumanyika and Lloyd 1996; Zar 1996). In 2009 and the 2010, the CCC was used as a measure of agreement in more than 60 medical publications in areas such as respiratory illness (Dixon, Sugar, Zinreich, Slavin, Corren, Naclerio, Ishii, Cohen, Brown, Wise and Irvin 2009; Kocks, Kerstjens, Snijders, de Vos, Biermann, van Hengel, Stribos, Bosveld and van der Molen 2010), sleep (Khawaja, Olson, van der Walt, Bukartyk, Somers, Dierkhising and Morgenthaler 2010), pediatrics (Liottol, Radaelli, Orsil, Taricco, Roggerol, Giann, Consonni, Moscal and Cetin 2010), neurology (MacDougall, Weber, McGarvie, Halmagyi and Curthoys 2009), and radiology (Mazaheri, Hricak, Fine, Akin, Shukla-Dave, Ishill, Moskowitz, Grater, Reuter, Zakian, Touijer and Koutcher 2009).

The kappa coefficient (κ) (Cohen 1960) and the weighted kappa coefficient (Cohen 1968) are the most popular indices for measuring agreement for discrete and ordinal outcomes, respectively (Fleiss 1981). Let Y_{i1} and Y_{i2} denote two binary random variables taking values 0 and 1 with probabilities denoted by $\pi_1 = \Pr(Y_{i1} = 1)$ and $\pi_2 = \Pr(Y_{i2} = 1)$. Kappa corrects the percentage of agreement between raters by taking into account the proportion of agreement expected by chance (calculated under independence), and is defined as follows:

$$\kappa = \frac{P_o - P_e}{1.0 - P_e}, \quad (2)$$

where P_e is the probability that the pair of binary responses are equal assuming independence ($\pi_1\pi_2 + (1-\pi_1)(1-\pi_2)$) and P_o is the probability that the pair are equal (Cohen 1960). The difference $P_o - P_e$ is the excess of agreement over chance agreement. A value of 0 for κ indicates no agreement beyond chance and a value of 1 indicates perfect agreement (Fleiss 1981). Disadvantages of kappa are that it is a function of the marginal distribution of the raters (Fleiss, Nee and Landis 1979; Tanner and Young 1985) and its range depends on the number of ratings per subject (Fleiss *et al.* 1979). Robieson (1999) noted that the CCC computed from ordinal scaled data is equivalent to the weighted kappa when integer scores are used. Kappa has been used to measure the validity and reproducibility of the similarity between twins (Klar, Lipsitz and Ibrahim 2000), different epidemiologic tools (Maclure and Willett 1987), and control-informant agreement from case-control studies (Korten, Jorm, Henderson, McCusker and Creasey 1992).

The value of sample surveys have been well recognized and estimation for data collected from sample surveys has been widely documented (Hansen, Hurwitz and Madow 1953; Cochran 1963; Kish 1965). For example, a number of federal studies conducted in the U.S. to obtain estimates of the health of the population are based on national surveys, such as the National Health Interview Survey (NHIS), the Behavioral Risk Factor Surveillance System (BRFSS), and the National Health and Nutrition Examination Surveys (NHANES). Each of these studies incorporates complex survey design structure, namely oversampling of subpopulations, stratification and clustering. These designs are often used to improve precision, provide estimates for subpopulations, or reduce costs associated with frame development. In order to draw design-based inference to the targeted population for complex survey designs, estimators and their variances include sampling weights and account for the design structure to obtain unbiased estimates. In addition, by including the sampling weights and incorporating the sample design in analyses, any potential correlation from the clusters in a multistage design is taken into account so that the standard errors of the estimators are not underestimated.

Often researchers are not interested in testing whether their estimation of agreement using either the CCC or kappa is significantly different from zero. Their interest is to report the confidence intervals along with their estimates (*e.g.*, Dixon *et al.* 2009; Mazaheri *et al.* 2009). Similar to the Pearson correlation coefficient, there is no target value that can be used to judge if agreement is strong. Therefore, it is essential that judgment of agreement between any test and reference methods should be made with an established degree of certainty. In some situations, studies are conducted that require hypothesis testing or comparisons of agreement indexes for more than one new methods against a reference method. For examples, Khawaja *et al.* (2010) tested the equality of two CCCs that compared the apnea hypopnea index (AHI) from the first 2 and 3 hours of sleep with the gold standard AHI from FN-PSG (FN-AHI). In radiology research, associations between volume measurements of prostate tumor from imaging and also from pathologic examination were assessed by comparing CCCs. The two imaging methods were tested for equality of agreement with the pathologic results (Mazaheri *et al.* 2009). Tests of equal kappa have been used to compare visual assessment and computerized planimetry in assessing cervical ectopy (Gilmour, Ellerbrock, Koulos, Chiasson, Williamson, Kuhn and Wright 1997; Williamson, Manatunga and Lipsitz 2000), and in comparing monozygotic and dizygotic twins in terms of cholesterol levels (Feinleib, Garrison, Fabsitz, Christian, Hrubec, Borhani, Kannel, Roseman, Schwartz and Wagner 1977).

As illustrated in the two NHANES III examples in Section 3, large differences can exist between the weighted and unweighted estimates of parameter estimate standard errors in survey studies. Failure to include sampling weights and take into account the sample design in analyses will result in underestimation of standard errors and incorrect inference. This is especially important for surveys repeated every few years, and researchers often have a special interest in comparing changes among domains or sub-populations. For instance, in the first NHANES III application, we compare the agreement between self reported and measured body weights at examination in adolescents. Computing accurate standard errors (confidence intervals) are necessary if interest is to compare the CCC across domains, such as normal weight and obese subgroups.

We provide weighted measures of the CCC and kappa coefficient, along with the variance estimators of these measures accounting for the sampling design. In Section 2, we present a generalized estimating equations approach for estimating these two agreement coefficients from sample survey data. In Section 3, we illustrate our method with data collected from the NHANES III study. We use body measurement data to estimate ρ_c for assessing the agreement between self-reported and actual weight. We also use oral health data to estimate κ for assessing the agreement between two definitions of periodontal disease. We account for stratification and clustering, and incorporate weights of the survey design in both examples. We conclude with a short discussion.

2. Methods

We propose a general approach for estimating the CCC and kappa from sample survey data using two GEE approaches. For the CCC, three sets of estimating equations are required. A first set of estimating equations models the distribution of the continuous responses. Following Barnhart and Williamson (2001), a second set of estimating equations is used to estimate the variances of the continuous responses. A third set of estimating equations estimates the CCC by modeling the covariance between the paired continuous responses and the estimates of the means and variances from the first two sets of estimating equations. For κ , only two sets of estimating equations are required. A first set of estimating equations models the marginal distribution of the binary responses. Following Lipsitz, Laird and Brennan (1994), a second set of estimating equations is introduced to estimate κ by modeling a binary random variable depicting agreement between two responses on a subject.

In order to account for variable selection probabilities, weight matrices are incorporated into each set of estimating

equations. Standard error estimation of the proposed $\hat{\rho}_c$ and $\hat{\kappa}$ from sample survey data are conducted with the Taylor series linearization method. We also show how standard error estimation of the proposed estimators can be accomplished by using the jackknife approach.

Assume a sample survey is conducted with stratification, clustering, and unequal probabilities of selection. Let Y_{hij} denote the response variable for the j^{th} member ($j = 1, \dots, m_{hi}$) of the i^{th} cluster ($i = 1, \dots, n_h$) of the h^{th} stratum ($h = 1, \dots, H$). Averaging over all possible samples, the corresponding expected value is $E[Y_{hij}] = \mu_{hij}$ if Y_{hij} is a continuous response, and the corresponding probability $E[Y_{hij}] = \Pr[Y_{hij} = 1] = \pi_{hij}$ if Y_{hij} is a binary response. The sampling weight w_{hij} is the inverse of the probability of selection for the j^{th} member of the i^{th} cluster of the h^{th} stratum.

2.1 The concordance correlation coefficient

Liang and Zeger (1986) developed moment-based methods for analyzing correlated observations from the same cluster (e.g., repeated measurements over time on the same individual or observations on multiple members of the same family). The GEE approach results in consistent marginal parameter estimation, even with misspecification of the correlation structure by using a robust “sandwich” estimator of variance. We use the GEE approach to analyze sample survey data by additionally incorporating a sampling weight matrix as follows:

$$\sum_{h=1}^H \sum_{i=1}^{n_h} \mathbf{D}'_{hi} \mathbf{W}_{hi} \mathbf{V}_{hi}^{-1} (\mathbf{Y}_{hi} - \boldsymbol{\mu}_{hi}(\hat{\boldsymbol{\mu}})) = \mathbf{0},$$

where \mathbf{D}'_{hi} is the $(q \times m_{hi})$ derivative matrix $d[\boldsymbol{\mu}_{hi}]'/d\boldsymbol{\mu}$, \mathbf{W}_{hi} is a $(m_{hi} \times m_{hi})$ main diagonal matrix consisting of the person-specific sampling weights w_{hij} , \mathbf{V}_{hi} is a $(m_{hi} \times m_{hi})$ working variance-covariance matrix for the within-cluster responses, \mathbf{Y}_{hi} is a $(m_{hi} \times 1)$ response vector consisting of the responses Y_{hij} , and $\boldsymbol{\mu}_{hi} = E[\mathbf{Y}_{hi}]$ is possibly a function of the $(q \times 1)$ parameter vector $\boldsymbol{\beta}$. The GEE can then be solved non-iteratively, resulting in the usual estimate

$$\hat{\boldsymbol{\mu}} = \left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} Y_{hij} \right) / \left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \right)$$

if we are estimating a common mean $\boldsymbol{\mu} = \boldsymbol{\beta}$ ($q = 1$) and are using an independence working covariance matrix.

Assume a pair of continuous responses are observed for the j^{th} member of the i^{th} cluster of the h^{th} stratum, Y_{hij1} and Y_{hij2} , and their expected values are μ_{hij1} and μ_{hij2} . Again, assume we are estimating common means μ_1 and μ_2 without covariates for the pair of within-subject continuous responses, which can be estimated by using the above generalized estimating equation.

Barnhart and Williamson (2001) demonstrated how three sets of generalized estimating equations can be used to model the CCC defined in (1) using correlated data. We extend Barnhart and Williamson's (2001) second set of GEE equations to estimate the variances of the continuous responses by again incorporating a weight matrix as follows:

$$v_2(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2) = \sum_{h=1}^H \sum_{i=1}^{n_h} \mathbf{F}'_{hi} \mathbf{W}_{hi} \mathbf{H}_{hi}^{-1} (\mathbf{Y}_{hi}^2 - \delta_{hi}^2(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2)) = 0,$$

where \mathbf{F}'_{hi} is the $(2 \times 2m_{hi})$ derivative matrix $d[\delta_{hi}^2]' / d\sigma^2$ with $\sigma^2 = [\sigma_1^2, \sigma_2^2]'$, \mathbf{W}_{hi} is a $(2m_{hi} \times 2m_{hi})$ main diagonal matrix consisting of the person-specific sampling weights w_{hij} , \mathbf{H}_{hi} is a $(2m_{hi} \times 2m_{hi})$ working variance-covariance matrix for the within-cluster squared responses, $\mathbf{Y}_{hi}^2 = [Y_{hi1}^2, Y_{hi12}^2, Y_{hi21}^2, Y_{hi22}^2, \dots, Y_{him_{hi}1}^2, Y_{him_{hi}2}^2]'$ is a $(2m_{hi} \times 1)$ response vector of the continuous variables, and $\delta_{hi}^2 = E[\mathbf{Y}_{hi}^2]$. Although δ_{hi}^2 is a function of both the variance terms σ_1^2 and σ_2^2 and the means μ_1 and μ_2 , it is assumed that the means are fixed in δ_{hi}^2 and one only takes derivatives of δ_{hi}^2 with respect to the variances. Again we choose the $(2m_{hi} \times 2m_{hi})$ matrix \mathbf{H}_{hi} to be the "independence" working variance-covariance matrix and the $(2m_{hi} \times 1)$ column vector $\delta_{hi}^2 = [\sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \dots, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2]'$ because we are assuming common variances and means across all strata and clusters. The above GEE can thus be solved non-iteratively:

$$\hat{\sigma}_p^2 = \left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{M_{hi}} W_{hijp} Y_{hijp}^2 \right) / \left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{M_{hi}} W_{hijp} \right) - \hat{\mu}_p^2,$$

for the p^{th} measurement in the pair, $p = 1, 2$.

The CCC can be estimated in a third set of estimating equations by using the pairwise products of the responses to model σ_{12} , once the means and variances are estimated. Let $\mathbf{U}_{hi} = [Y_{hi1}Y_{hi12}, Y_{hi21}Y_{hi22}, \dots, Y_{him_{hi}1}Y_{him_{hi}2}]'$ be a $(m_{hi} \times 1)$ vector of pairwise products of the responses and denote $\theta_{hi} = E[\mathbf{U}_{hi}]$, which is a function of the means, variances, and CCC. We solve for $\hat{\rho}_c$ in a third set of estimating equations:

$$v_3(\hat{\rho}_c, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2) = \sum_{h=1}^H \sum_{i=1}^{n_h} \mathbf{C}'_{hi} \mathbf{W}_{hi} \mathbf{K}_{hi}^{-1} (\mathbf{U}_{hi} - \theta_{hi}(\hat{\rho}_c, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2)) = 0,$$

where \mathbf{C}'_{hi} is a $(1 \times m_{hi})$ derivative vector $= \partial \theta_{hi} / \partial \rho_c$, \mathbf{W}_{hi} is a $(m_{hi} \times m_{hi})$ main diagonal matrix consisting of the person-specific sampling weights w_{hij} , and \mathbf{K}_{hi} is a $(m_{hi} \times m_{hi})$ working covariance matrix that we choose to be the "independence" covariance matrix. The above GEE can be solved non-iteratively:

$$\hat{\rho}_c = \frac{2\hat{\sigma}_{12}}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + (\hat{\mu}_1 - \hat{\mu}_2)^2},$$

where

$$\hat{\sigma}_{12} = \frac{\left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{M_{hi}} W_{hij12} Y_{hij1} Y_{hij2} \right)}{\left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{M_{hi}} W_{hij12} \right)} - \hat{\mu}_1 \hat{\mu}_2.$$

2.2 Linearization estimator of variance

The usual robust estimators of variance for the means and CCC from the GEE approach are invalid here because they do not take into account the sampling structure, only the correlation of observations made on the same individual. We propose standard error estimation using the Taylor series linearization method (Binder 1983; Binder 1996). The first derivatives of ρ_c (equation 1) with respect to μ_1 , μ_2 , σ_1^2 , σ_2^2 , and σ_{12} are:

$$\begin{aligned} \frac{\partial \rho_c}{\partial \mu_1} &= \frac{-4\sigma_{12}(\mu_1 - \mu_2)}{D^2}, \\ \frac{\partial \rho_c}{\partial \mu_2} &= \frac{-4\sigma_{12}(\mu_2 - \mu_1)}{D^2}, \\ \frac{\partial \rho_c}{\partial \sigma_1^2} &= \frac{-2\sigma_{12}}{D^2}, \\ \frac{\partial \rho_c}{\partial \sigma_2^2} &= \frac{-2\sigma_{12}}{D^2}, \\ \frac{\partial \rho_c}{\partial \sigma_{12}} &= \frac{2}{D}, \end{aligned}$$

where $D = \sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2$. Thus

$$\begin{aligned} \hat{\rho}_c - \rho_c &\approx \left(\frac{\partial \rho_c}{\partial \mu_1} \right) (\hat{\mu}_1 - \mu_1) + \left(\frac{\partial \rho_c}{\partial \mu_2} \right) (\hat{\mu}_2 - \mu_2) \\ &+ \left(\frac{\partial \rho_c}{\partial \sigma_1^2} \right) (\hat{\sigma}_1^2 - \sigma_1^2) \\ &+ \left(\frac{\partial \rho_c}{\partial \sigma_2^2} \right) (\hat{\sigma}_2^2 - \sigma_2^2) + \left(\frac{\partial \rho_c}{\partial \sigma_{12}} \right) (\hat{\sigma}_{12} - \sigma_{12}) \\ &= \frac{-4\sigma_{12}(\mu_1 - \mu_2)}{D^2} (\hat{\mu}_1 - \mu_1) \\ &+ \frac{-4\sigma_{12}(\mu_2 - \mu_1)}{D^2} (\hat{\mu}_2 - \mu_2) \\ &+ \frac{-2\sigma_{12}}{D^2} (\hat{\sigma}_1^2 - \sigma_1^2) + \frac{-2\sigma_{12}}{D^2} (\hat{\sigma}_2^2 - \sigma_2^2) \\ &+ \frac{2}{D} (\hat{\sigma}_{12} - \sigma_{12}). \end{aligned}$$

The above equation can be rearranged into two parts, one involving the parameter estimates $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2$, and $\hat{\sigma}_{12}$ and the other involving only parameters which does not

contribute to the variance estimation of $\hat{\rho}_c$. Thus the first part becomes

$$\begin{aligned} & -\frac{4\sigma_{12}(\mu_1 - \mu_2)}{D^2}\hat{\mu}_1 - \frac{4\sigma_{12}(\mu_2 - \mu_1)}{D^2}\hat{\mu}_2 \\ & - \frac{2\sigma_{12}}{D^2}\hat{\sigma}_1^2 - \frac{2\sigma_{12}}{D^2}\hat{\sigma}_2^2 + \frac{2}{D}\hat{\sigma}_{12} \\ & = -\frac{2\sigma_{12}}{D^2}(2(\mu_1 - \mu_2)(\hat{\mu}_1 - \hat{\mu}_2) + \hat{\sigma}_1^2 + \hat{\sigma}_2^2) + \frac{2}{D}\hat{\sigma}_{12} \\ & = -\frac{2\sigma_{12}}{D^2}\left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} 2(\mu_1 - \mu_2)(w_{hij}^* Y_{hij1} - w_{hij}^* Y_{hij2}) \right. \\ & \quad \left. + w_{hij}^* (Y_{hij1} - \mu_1)^2 + w_{hij}^* (Y_{hij2} - \mu_2)^2\right) \\ & \quad + \frac{2}{D} \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}^* (Y_{hij1} - \mu_1)(Y_{hij2} - \mu_2) \end{aligned} \quad (3)$$

where $w_{hij}^* = w_{hij} / (\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij})$. Equation (3) becomes a linear function of the data after the summation is moved to the front, which we can then express as $\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}^* z_{hij}$, where

$$\begin{aligned} z_{hij} = & -\frac{2\sigma_{12}}{D^2}(2(\mu_1 - \mu_2)(Y_{hij1} - Y_{hij2}) \\ & + (Y_{hij1} - \mu_1)^2 + (Y_{hij2} - \mu_2)^2) \\ & + \frac{2}{D}(Y_{hij1} - \mu_1)(Y_{hij2} - \mu_2). \end{aligned} \quad (4)$$

One then creates a random variable \hat{z}_{hij} based on equation (4) that replaces the parameters with their respective estimates. The variance of this new estimator \hat{z}_{hij} is an approximation for the variance of $\hat{\rho}_c$, which can be estimated using standard survey software (see Appendix).

2.3 Jackknife estimator of variance

We also use the jackknife technique for standard error estimation of the parameters following Rust and Rao (1996, Section 2.1) for comparison with the linearization estimates. The jackknife technique is implemented by calculating a set of replicate estimates and estimating the variance using them. A replicate data set is created for each cluster by deleting all observations from the given cluster from the sample. The weights of all other observations in the stratum containing the cluster are inflated by a factor $n_h / (n_h - 1)$. Weights in the other strata remain unchanged. Thus, the new weights for the replicated data set created by removing cluster i from stratum h are:

$$\begin{aligned} \omega_{kij}^{(hi)} &= w_{kij} && \text{if } k \neq h \text{ (different strata)} \\ \omega_{hij}^{(hi)} &= w_{hij} n_h / (n_h - 1) && \text{if } l \neq i \\ &&& \text{(same strata but different clusters)} \\ \omega_{hij}^{(hi)} &= 0 && \text{(for the cluster being removed).} \end{aligned}$$

The resulting jackknife variance estimator for $\hat{\rho}_c$ is

$$v_J(\hat{\rho}_c) = \sum_{h=1}^H \left(\frac{n_h - 1}{n_h} \right) \sum_{i=1}^{n_h} (\hat{\rho}_{c(hi)} - \hat{\rho}_c)^2$$

where $\hat{\rho}_{c(hi)}$ is estimated in the same way as $\hat{\rho}_c$, but using the recalculated weights $\omega^{(hi)}$ instead of the weights ω . The jackknife estimators for the means are similarly calculated.

2.4 The kappa coefficient

Assume a pair of binary responses are observed for the j^{th} member of the i^{th} cluster of the h^{th} stratum, Y_{hij1} and Y_{hij2} , and their expected values are the probabilities π_{hij1} and π_{hij2} . Again assume we are estimating common probabilities π_1 and π_2 without covariates for the pair of within-subject binary responses. Lipsitz *et al.* (1994) demonstrated how two sets of generalized estimating equations can be used to develop simple non-iterative estimates of the κ -coefficient that can be used for unbalanced data as previous estimates of kappa and its variance were only proposed for balanced data. They defined the binary random variable $U_{hij} = Y_{hij1} Y_{hij2} + (1 - Y_{hij1})(1 - Y_{hij2}) = 1$ if both responses in the pair agree and 0 otherwise. Accordingly, $E[U_{hij}] = P_o$, which denotes the probability of observed agreement and is assumed here to be constant over all strata, clusters, and pairs of observations. Now let $E[Y_{hij1} Y_{hij2}] = \Pr[Y_{hij1} = Y_{hij2} = 1] = \omega$. The probability of observed agreement can be expressed as $P_o = 1 - \pi_1 - \pi_2 + 2\omega$. The probability of expected agreement by chance is defined as $P_e = \pi_1 \pi_2 + (1 - \pi_1)(1 - \pi_2)$ and is estimated by $\hat{P}_e = \hat{\pi}_1 \hat{\pi}_2 + (1 - \hat{\pi}_1)(1 - \hat{\pi}_2)$, where $\hat{\pi}_1$ and $\hat{\pi}_2$ are calculated in the first set of estimating equations.

We can derive estimates of κ from sample survey data following the approach for the CCC in Section 2.1. We can incorporate the survey weight matrices into Lipsitz *et al.*'s (1994) two sets of GEE equations for estimating kappa. Then, by choosing "independence" working covariance matrices for the two sets of equations as in Lipsitz *et al.*'s (1994) approach, we arrive at the following non-iterative estimate of kappa for sample survey data:

$$\hat{\kappa} = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} U_{hij} - \hat{P}_e \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}}{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} - \hat{P}_e \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}}. \quad (5)$$

This estimator is identical to Lumley's (2010), which can be computed using the R software survey package and svykappa function.

Standard error estimation of $\hat{\kappa}$ can be conducted similarly to that of $\hat{\rho}_c$ using the Taylor series linearization method. The first derivatives of kappa with respect to P_o, π_1 , and π_2 are:

$$\begin{aligned}\frac{\partial \kappa}{\partial P_o} &= \frac{1}{1 - P_e}, \\ \frac{\partial \kappa}{\partial \pi_1} &= \frac{(1 - P_o)(1 - 2\pi_2)}{(1 - P_e)^2}, \\ \frac{\partial \kappa}{\partial \pi_2} &= \frac{(1 - P_o)(1 - 2\pi_1)}{(1 - P_e)^2}.\end{aligned}$$

Thus

$$\begin{aligned}\hat{\kappa} - \kappa &\approx \left(\frac{\partial \kappa}{\partial P_o} \right) (\hat{P}_o - P_o) \\ &\quad + \left(\frac{\partial \kappa}{\partial \pi_1} \right) (\hat{\pi}_1 - \pi_1) + \left(\frac{\partial \kappa}{\partial \pi_2} \right) (\hat{\pi}_2 - \pi_2) \\ &= \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}^* z_{hij},\end{aligned}$$

where $w_{hij}^* = w_{hij} / (\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij})$ and

$$\begin{aligned}z_{hij} &= \left(\frac{\partial \kappa}{\partial P_o} \right) U_{hij} + \left(\frac{\partial \kappa}{\partial \pi_1} \right) Y_{hij1} + \left(\frac{\partial \kappa}{\partial \pi_2} \right) Y_{hij2} \\ &= \frac{U_{hij}}{1 - P_e} + \frac{(1 - P_o)[Y_{hij1}(1 - 2\pi_2) + Y_{hij2}(1 - 2\pi_1)]}{(1 - P_e)^2}.\end{aligned}\quad (6)$$

Replacing the parameters in (6) with their respective estimates, one then treats \hat{z}_{hij} as a random variable and estimates its variance using standard survey software that accounts for the sampling design. The variance of this new estimator \hat{z}_{hij} is an approximation for the variance of $\hat{\kappa}$. The jackknife method can also be used to estimate the variance of $\hat{\kappa}$.

3. NHANES III survey

We used data from the Third National Health and Nutrition Examination Survey to illustrate our method. NHANES III was conducted by the National Center for Health Statistics of the Centers for Disease Control and Prevention and was designed as a six-year survey divided into two phases (1988-1991 and 1991-1994). The data were collected using a complex, multistage, probability sampling design to select participants representative of the civilian, non-institutionalized US population. Details of the survey design and analytic and reporting guidelines were published in the NHANES III reference manuals and reports (National Center for Health Statistics 1996).

3.1 The adolescent weight study

Obesity is a rapidly increasing public health problem with surveillance most often based on self-reported values of height and weight. A series of recent studies and systemic

reviews have attempted to assess the agreement between self-reported and measured weight, especially in the adolescent population. The general findings suggest that self-reported weight was slightly lower than measured weight, and that a significant number of normal weight adolescents misperceive themselves as overweight and are engaging in unhealthy weight control behaviors (Field, Aneja and Rosner 2007; Gorber, Tremblay, Moher and Gorber 2007; Sherry, Jefferds and Grummer-Strawn 2007). Therefore, researchers have suggested that obesity prevention programs should address weight misperceptions and the harmful effects of unhealthy weight control methods even among normal weight adolescents (Talamayan, Springer, Kelder, Gorospe and Joye 2006). A similar Canadian study from the 2005 Canadian Community Health Survey that focused on adult individuals also showed that associations between obesity and health conditions may be overestimated if self-reported weight is used (Shield, Gorber and Tremblay 2008). We use data obtained from the Body Measurements (Anthropometry) component of the NHANES III study to estimate the CCC that measures agreement between self-reported and measured weight (pounds) obtained from adolescents (aged 12 through 16 years).

The self-reported weight was obtained just prior to the actual measurement of weight. We use data from the entire six-year survey period (both 1988-1991 and 1991-1994). For simplicity, we excluded one stratum which only had one PSU. Hence, there were 48 strata and each stratum had two PSUs. The sample weight labeled *wtpfex6* accounting for the differential selection probability was used in our analyses. There were 1,651 subjects with complete data for both weight measurements. The estimates of the self-reported and actual weights (in pounds) were 135.5 (s.e. = 1.8) and 136.3 (s.e. = 1.8), respectively, calculated using PROC SURVEYMEANS in SAS. The estimates of the standard errors based on the jackknife approach are the same as above.

The CCC is a natural choice for assessing the agreement between the two weight measurements because they are measured on the same scale and their ranges are similar (self-reported weight: 78 lbs ~ 350 lbs and actual weight: 73 lbs ~ 372 lbs) (Lin and Chinchilli 1997). The estimate of the CCC for measuring the agreement between the two definitions of weight using the proposed method is 0.93. The standard error of the estimate is 0.021 using the Taylor series linearization method. The jackknife standard error of 0.021 agrees closely with the linearization standard error. These statistics are summarized in Table 1 along with their values computed when the sampling structure is ignored. The standard errors for the estimates incorporating the sampling structure are much larger than the unweighted estimates.

Table 1
Unweighted and weighted average, CCC, and respective standard errors for adolescent self-reported and actual weight in pounds

	Self-reported	Actual	CCC
Unweighted Estimate	135.31	136.96	0.890
SE	0.76	0.80	0.0005
Weighted Estimate	135.47	136.30	0.926
SE	1.75	1.82	0.0205

Similar to the CCC, the usual Pearson correlation coefficient between the self-reported and the actual weight measures is also 0.93. In this case, the mean difference between the two weight measurements is just less than one pound. When subpopulations are examined, differences are noted in the CCC and the Pearson correlation coefficient. Consider a subpopulation of those individuals that had a measured weight > 200 lbs at examination. Summarizing the data for this subpopulation, the self-reported weight is on average 8 pounds less than the measured weight (223.2 lbs vs 231.4 lbs). There is a slight departure of the CCC (0.72) from the Pearson correlation coefficient (0.76). The discrepancy between the two measures increases in the more obese subgroup. In the subpopulation where measured weight is > 220 lbs, the means of self-reported and measured weights are 231.9 lbs and 248.8 lbs, respectively. The CCC is 0.67, whereas the Pearson correlation coefficient is 0.85. In this situation, the CCC reflects both the reproducibility and differences between the self-reported and measured means. Therefore, the CCC is informative and advantageous when considering these comparisons, particularly in domain analysis within a complex survey.

3.2 The oral health study

Slade and Beck (1999) used extent of pocket depth and loss of attachment as indices of periodontal conditions. Prevalence of periodontal disease using previously reported thresholds of pocket depth ≥ 4 mm and attachment loss ≥ 3 mm were estimated by Slade and Beck (1999, Table 1). Pocket depth may be reflective of inflammation rather than chronic periodontal disease and, thus, attachment level may be a more meaningful measure of periodontal destruction. However, pocket depth remains the recommended measurement in clinical practice (Winn, Johnson and Kingman 1999). Therefore, we compare the agreement of these two definitions of periodontal disease using the kappa coefficient.

We use the sample that was analyzed by Slade and Beck (1999). The data include 14,415 persons aged 13 or older who had complete pocket depth and attachment loss assessment by six designated dentists. We again use data from the entire six-year survey period (both 1988-1991 and 1991-1994). There were a total of 49 strata and each stratum

had two PSUs. The variable labeled sample weight, wtpfex6, accounting for differential selection probability, was used in our analyses.

The first definition of periodontal disease is pocket depth ≥ 4 mm and the second is maximum attachment loss ≥ 3 mm. For both variables we are using the maximum values among all teeth in an individual's mouth. The probability estimates of the attachment loss and pocket depth variables are 0.358 (jackknife s.e. = 0.0088) and 0.212 (jackknife s.e. = 0.016), respectively, using the proposed method. The asymptotic standard errors based on the usual Taylor series expansion (Woodruff 1971, produced by PROC SURVEYFREQ in SAS, version 9.1) are 0.0088 and 0.015, respectively.

Kappa is a natural choice for assessing the agreement between two binary ratings as it corrects for chance agreement (Fleiss 1981). The estimate of kappa for measuring the agreement between the two definitions of periodontal disease (pocket depth of ≥ 4 mm and attachment loss of ≥ 3 mm) using the proposed method is 0.307. The standard error of 0.0158 was obtained by both the Taylor series linearization and jackknife methods. Table 2 compares these results to the measures when the complex sampling structure is ignored. The standard error of the kappa coefficient is larger when accounting for the survey structure.

Table 2
Unweighted and weighted average, kappa, and respective standard errors for attachment loss and pocket depth

	Attachment Loss	Pocket Depth	Kappa
Unweighted Estimate	0.393	0.283	0.334
SE	0.004	0.004	0.008
Weighted Estimate	0.358	0.212	0.307
SE	0.009	0.016	0.0158

4. Discussion

The CCC and kappa evaluate the agreement between two measurements for continuous and categorical responses, respectively. In this paper, we have proposed a generalized estimating equation approach for estimating the CCC for a pair of continuous variables, and kappa for a pair of binary variables, from sample survey data where the data have been collected using complex survey features such as stratification or clustering. The usual sandwich estimator of the variance only accounts for repeated measurements made on the same individual, and does not account for the sampling framework (e.g., clustering, stratification, and weighting). In the GEE approach, standard error estimation of the estimators is conducted with the Taylor series linearization and jackknife approaches. If the data are not collected using complex survey features, the proposed estimators will be identical to the usual estimators. As is

evident in the two examples from the NHANES III study, we have shown the need to incorporate sampling weights and the sampling design features so that the standard errors are not underestimated when data are collected from a complex sampling design. Tables 1 and 2 show that there were large differences in the standard errors between weighted and unweighted estimates of the standard errors for both CCC and kappa. Confidence intervals that incorporate weights and the design features will allow correct inference.

In the appendix, we show steps for calculating the weighted measures of the CCC and kappa, along with their standard errors using standard survey software that incorporates the sampling weights, clustering and stratification. The GEE approach is advantageous because it is a convenient framework for developing estimators of the agreement coefficients and is easily extended to multiple raters, multiple methods, covariate adjustment and unbalanced cluster sizes. This design-based approach results in correct standard error estimation without assuming an underlying model and accounting for the sampling structure. If one is interested in estimating the agreement between two ordinal variables with kappa then Williamson *et al.*'s (2000) generalized estimating equation approach can be extended similarly to the proposed method.

Acknowledgements

We thank the anonymous reviewers and editor for their helpful comments. In particular, the editor gave extremely valuable suggestions for the introduction section.

Appendix

Steps for calculating the CCC and its standard error using standard survey software

- Step 1: Calculate the means of the continuous variables Y_{hij1} and Y_{hij2} using software for survey data that incorporates stratification, clustering, and sample weighting (e.g., PROC SURVEYMEANS in SAS).
- Step 2: Square the centered Y_{hij1} and Y_{hij2} values around their respective means.
- Step 3: Calculate the means of the squared centered Y_{hij1} and Y_{hij2} values using standard software for survey data. These means are the variance estimates of Y_{hij1} and Y_{hij2} . Calculate the mean of the product of the centered Y_{hij1} and Y_{hij2} values using standard software for survey data. This mean is the estimated covariance of Y_{hij1} and Y_{hij2} .
- Step 4: Calculate the CCC by substituting the estimated means and variances into equation (1). Create the new variable Z_{hij} based on equation (4).

- Step 5: Calculate the standard error of Z_{hij} using standard software for survey data. The standard error of Z_{hij} estimates the standard error of $\hat{\rho}_c$.

SAS CODE:

Let $y1$ and $y2$ denote the variables for the pair of continuous responses, and s , c and w denote the variables for strata, cluster and weight:

```
PROC SURVEYMEANS DATA=dataset MEAN; /* Step 1 above */;
  STRATA s;
  CLUSTER c;
  WEIGHT w;
  VAR y1 y2;
  ODS OUTPUT STATISTICS=stat;
data _null_;
  set stat (where=(varname='y1' ));
  call symputx('muy1', mean);
data _null_;
  set stat (where=(varname='y2' ));
  call symputx('muy2', mean);
data dataset; set dataset; /* Step 2 above */;
  cy1 = y1 - &muy1;
  cy2 = y2 - &muy2;
  vary1 = cy1 **2;
  vary2 = cy2 **2;
  covy12 = cy1 * cy2;
PROC SURVEYMEANS MEAN; /* Step 3 above */;
  STRATA s;
  CLUSTER c;
  WEIGHT w;
  VAR vary1 vary2 covy12;
  ODS OUTPUT STATISTICS=stat;
run;
data _null_;
  set stat (where=(varname='vary1' ));
  call symputx('vary1', mean);
data _null_;
  set stat (where=(varname='vary2' ));
  call symputx('vary2', mean);
data _null_;
  set stat (where=(varname='covy12' ));
  call symputx('covy12', mean);
data dataset; set dataset; /* Step 4 above */;
  d = &vary1 + &vary2 + (&muy1 - &muy2) ** 2;
  CCC = 2 * &covy12 / d;
  z = (2 / d) * (cy1 * cy2) - (2 * &covy12 / d / d) * ((cy1 **2) +
  (cy2 **2) + 2 * (&muy1 - &muy2) * (y1 - y2));
PROC SURVEYMEANS MEAN; /* Step 5 above */;
  STRATA s;
  CLUSTER c;
  WEIGHT w;
  VAR CCC z;
run;
```

Steps for calculating kappa and its standard error using standard survey software

Step 1: Estimate the probabilities of the binary variables Y_{hij1} and Y_{hij2} using software for survey data that incorporates stratification, clustering, and sample weighting (e.g., PROC SURVEYFREQ in SAS).

Step 2: Estimate $P_c (= \hat{\pi}_1\hat{\pi}_2 + (1 - \hat{\pi}_1)(1 - \hat{\pi}_2))$.

Step 3: Create the new agreement variable $U_{hij} (= Y_{hij1}Y_{hij2} + (1 - Y_{hij1})(1 - Y_{hij2}))$.

Step 4: Calculate the sum of the sample survey weights and the sum of the weighted U_{hij} (e.g., using PROC SURVEYMEANS in SAS). Estimate kappa using equation (2).

Step 5: Create a new variable z_{hij} using equation (6).

Step 6: Calculate the standard error of z_{hij} using standard software for survey data. The standard error of z_{hij} estimates the standard error of $\hat{\kappa}$.

References

- Barnhart, H.X., and Williamson, J.M. (2001). Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics*, 57, 931-940.
- Behavioral Risk Factor Surveillance System (BRFSS). <http://www.cdc.gov/BRFSS>.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 20, 37-46.
- Binder, D.A. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.
- Carrasco, J.L., and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, 59, 849-858.
- Chinchilli, V.M., Martel, J.K., Kumanyika, S. and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, 52, 341-353.
- Cochran, W.G. (1963). *Sampling Techniques*, 2nd Ed. New York: John Wiley & Sons, Inc.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.
- Dixon, A.E., Sugar, E.A., Zinreich, S.J., Slavin, R.G., Corren, J., Naclerio, R.M., Ishii, M., Cohen, R.I., Brown, E.D., Wise, R.A. and Irvin, C.G. (2009). Criteria to screen for chronic sinonasal disease. *Chest*, 136 (5), 1324-1332.
- Feinleib, M., Garrison, R.J., Fabsitz, R.R., Christian, J.C., Hrubec, Z., Borhani, N.O., Kannel, W.B., Roseman, R., Schwartz, J.T. and Wagner, J.O. (1977). The NHLBI Twin Study of cardiovascular disease risk factors: Methodology and summary of results. *American Journal of Epidemiology*, 106, 284-295.
- Field, A.E., Aneja, P. and Rosner, B. (2007). The validity of self-reported weight change among adolescents and young adults. *Obesity*, 15, 2357-2364.
- Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Edition. New York: John Wiley & Sons, Inc.
- Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, Inc.
- Fleiss, J.L., Nee, J.C.M. and Landis, J.R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86, 974-977.
- Gilmour, E., Ellerbrock, T.V., Koulos, J.P., Chiasson, M.A., Williamson, J.M., Kuhn, L. and Wright, T.C. (1997). Measuring cervical ectopy: Direct visual assessment versus computerized planimetry. *American Journal of Obstetrics and Gynecology*, 176, 108-111.
- Gorber, S.C., Tremblay, M., Moher, D. and Gorber, B. (2007). A comparison of direct vs. self-report measures for assessing height, weight and body mass index: A systematic review. *Obesity Review*, 8, 373-374.
- Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons, Inc. Vols I and II.
- Khawaja, I.S., Olson, E.J., van der Walt, C., Bukartyk, J., Somers, V., Dierkhising, R. and Morgenthaler, T.I. (2010). Diagnostic accuracy of split-night polysomnograms. *Journal of Clinical Sleep Medicine*, 6 (4), 357-362.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Klar, N., Lipsitz, S.R. and Ibrahim, J.G. (2000). An estimating equations approach for modeling kappa. *Biometrical Journal*, 42, 45-58.
- Kocks, J.W., Kerstjens, H.A., Snijders, S.L., de Vos, B., Biermann, J.J., van Hengel, P., Stribos, J.H., Bosveld, H.E. and van der Molen, T. (2010). Health status in routine clinical practice: validity of the clinical COPD questionnaire at the individual patient level. *Health and Quality of Life Outcomes*, 8, 135-141.
- Korten, A.E., Jorm, A.F., Henderson, A.S., McCusker, E. and Creasey, H. (1992). Control-informant agreement on exposure history in case-control studies of Alzheimer's disease. *International Journal of Epidemiology*, 21, 1121-1131.
- Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255-268.
- Lin, L. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, 48, 599-604.

- Lin, L., and Chinchilli, V. (1997). Rejoinder to the letter to the editor from Atkinson and Nevill. *Biometrics*, 53, 777-778.
- Liottol, N., Radaelli, T., Orsi, A., Taricco, E., Roggerol, P., Giann, M.L., Consonni, D., Mosca, F. and Cetin, I. (2010). Relationship between in utero sonographic evaluation and subcutaneous plicometry after birth in infants with intrauterine growth restriction: An exploratory study. *Italian Journal of Pediatrics*, 36, 70-77.
- Lipsitz, S.R., Laird, N.M. and Brennan, T.A. (1994). Simple moment estimates of the κ -coefficient and its variance. *Applied Statistics*, 43, 309-323.
- Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- MacDougall, H.G., Weber, K.P., McGarvie, L.A., Halmagyi, G.M. and Curthoys, I.S. (2009). The video head impulse test. Diagnostic accuracy in peripheral vestibulopathy. *Neurology*, 73, 1134-1141.
- Maclure, M., and Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126, 161-169.
- Mazaheri, Y., Hricak, H., Fine, S.W., Akin, O., Shukla-Dave, A., Ishill, N.M., Moskowitz, C.S., Grater, J.E., Reuter, V.E., Zakian, K.L., Touijer, K.A. and Koutcher, J.A. (2009). Prostate tumor volume measurement with combined T2-weighted imaging and diffusion-weighted MR: Correlation with pathologic tumor volume. *Radiology*, 252 (2), 449-457.
- National Center for Health Statistics (2011). Third National Health and Nutrition Examination Survey, 1988-1994, NHANES III Examination data file (CD-ROM). <http://www.cdc.gov/nchs/nhanes.htm>.
- National Health Interview Survey (NHIS) (2011). <http://www.cdc.gov/nchs/nhis.htm>.
- Robieson, W. (1999). On weighted kappa and concordance correlation coefficient. Ph.D. thesis, University of Illinois in Chicago/Graduate College/Mathematics.
- Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Sherry, B., Jefferds, M.E. and Grummer-Strawn, L.M. (2007). Accuracy of adolescent self-report of height and weight in assessing overweight status: A literature review. *Archives of Pediatrics Adolescent Medicine*, 161, 1154-1161.
- Shield, M., Gorber, S.C. and Tremblay, M.S. (2008). Effects of measurement on obesity and morbidity. *Health Reports*, 19, 77-84.
- Slade, G.D., and Beck, J.D. (1999). Plausibility of periodontal disease estimates from NHANES III. *Journal of Public Health Dentistry*, 59, 67-72.
- Talamayan, K.S., Springer, A.E., Kelder, S.H., Gorospe, E.C. and Joye, K.A. (2006). Prevalence of overweight misperception and weight control behaviors among normal weight adolescents in the United States. *The Scientific World Journal*, 6, 365-373.
- Tanner, M.A., and Young, M.A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, 80, 175-180.
- Williamson, J.M., Manatunga, A.K. and Lipsitz, S.R. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics*, 1, 191-202.
- Winn, D.M., Johnson, C.L. and Kingman, A. (1999). Periodontal disease estimates in NHANES III: Clinical measurement and complex sample design issues. *Journal of Public Health Dentistry*, 59, 73-78.
- Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.
- Zar, J.H. (1996). *Biostatistical Analysis*. Upper Saddle River, New Jersey: Prentice Hall International.

Combining synthetic data with subsampling to create public use microdata files for large scale surveys

Jörg Drechsler and Jerome P. Reiter¹

Abstract

To create public use files from large scale surveys, statistical agencies sometimes release random subsamples of the original records. Random subsampling reduces file sizes for secondary data analysts and reduces risks of unintended disclosures of survey participants' confidential information. However, subsampling does not eliminate risks, so that alteration of the data is needed before dissemination. We propose to create disclosure-protected subsamples from large scale surveys based on multiple imputation. The idea is to replace identifying or sensitive values in the original sample with draws from statistical models, and release subsamples of the disclosure-protected data. We present methods for making inferences with the multiple synthetic subsamples.

Key Words: Confidentiality; Disclosure; Multiple imputation.

1. Introduction

National Statistical Institutes (NSIs) like the U.S. Census Bureau and Statistics Canada conduct large scale surveys that are highly valued by secondary data analysts, such as the American Community Survey (ACS) and the National Longitudinal Survey of Children and Youth (NLSCY). While these analysts desire access to as much data as possible, the NSI also must protect the confidentiality of survey participants' identities and sensitive attributes. A common strategy for reducing disclosure risks in large scale studies is to release subsamples of the original survey data; for example, the Census Bureau releases a subsample from the collected ACS data comprising 1% of all U.S. households (the collected ACS data comprise 2.5% of all households), and Statistics Canada releases a 20% sample of individuals from the NLSCY. See Willenborg and de Waal (2001) and Reiter (2005) for discussions of the confidentiality protection engendered by sampling. Typically, however, subsampling alone does not eliminate disclosure risks, particularly for units in the subsample with unusual combinations of characteristics. NSIs therefore alter data before dissemination. For example, in the ACS, the Census Bureau performs data swapping, topcoding of selected variables, aggregating of geography, and age perturbation; in the NLSCY, Statistics Canada uses data swapping and suppression.

When implemented with high intensity, as may be necessary to protect confidentiality in highly visible surveys, standard disclosure limitation strategies can seriously distort inferences (Winkler 2007; Elliott and Purdam 2007; Drechsler and Reiter 2010). Further, for many standard techniques it is difficult for data analysts - especially those

without advanced statistical training - to properly account for the effects of the disclosure control in estimation. Motivated by these limitations, we propose a new approach for generating public use microdata samples from large scale surveys called subsampling with synthesis. The basic idea is to replace identifying or sensitive values in the original sample with multiple draws from statistical models estimated with the original data file, and release subsamples of the disclosure-protected data. The subsamples can comprise one common set of records, or they can be taken independently.

This approach is a variant of partially synthetic data (Little 1993; Reiter 2003), which has been used in the U.S. to create several public use data products, including the Survey of Income and Program Participation, the Longitudinal Business Database, the Survey of Consumer Finances, the American Community Survey group quarters data, and OnTheMap. The approach proposed here differs from partial synthesis because of the subsampling, which necessitates adjustments to the inferential methods of Reiter (2003); these are presented here. The approach also differs from the methods for creating synthetic public use microdata samples of census data developed recently by Drechsler and Reiter (2010). In subsampling with synthesis, the initial data come from a survey and not from a census; thus, inferences must account for the additional uncertainty that results from the initial sampling.

2. General approach

We now describe the data generation and inferential procedures for the two approaches to subsampling with

1. Jörg Drechsler, Institute for Employment Research, Department for Statistical Methods, Regensburger Straße 104, 90478 Nürnberg, Germany. E-Mail: joerg.drechsler@iab.de; Jerome P. Reiter, Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251. E-Mail: jerry@stat.duke.edu.

synthesis: releasing different (independent) subsamples, and releasing a common set of records in each subsample. The data generation methods, as well as methods for making valid inferences from the multiple datasets, depend on the subsampling approach. For both approaches, we let D denote the original survey data of n_1 units sampled from a population consisting of N units. We initially assume that the original sampling design is a simple random sample; we later extend to stratified sampling. We assume that all sampled units fully respond in D . Unlike for standard partial synthesis (Reiter 2004), methods have not been developed to handle missing data and synthesis with subsampling simultaneously. We focus here on general descriptions of the approaches and presentation of the inferential methods. We do not discuss synthesis model building strategies; see Drechsler and Reiter (2009) and the references therein for guidance.

2.1 Releasing different random subsamples

2.1.1 Summary of approach

To begin, the NSI creates m partially synthetic datasets, $D_{syn} = \{D_i: i = 1, \dots, m\}$, for the original survey following the approach of Reiter (2003). Specifically, the NSI replaces identifying or sensitive values in D with multiple imputations. Synthesis models are estimated using only the records whose values will be synthesized. The synthesis is done independently m times, resulting in D_{syn} . The NSI then takes a simple random subsample of $n_2 < n_1$ records from each D_i . These m subsamples, $d_{syn} = \{d_i: i = 1, \dots, m\}$, are released to the public.

The analyst of d_{syn} seeks inferences about some estimand Q , such as a population mean or regression coefficient. In each d_i , the analyst estimates Q with some point estimator q and estimates the variance of q with some estimator u , where the analyst specifies q and u acting as if d_i were the collected data. Here, u is specified ignoring any finite population correction factors; for example, when q is the sample mean, $u = s^2/n_2$, with s^2 being the sample variance. For $i = 1, \dots, m$, let q_i and u_i be the values of q and u in d_i . The following quantities are needed for inferences.

$$\bar{q}_m = \sum_{i=1}^m q_i / m \quad (1)$$

$$b_m = \sum_{i=1}^m (q_i - \bar{q}_m)^2 / (m-1) \quad (2)$$

$$\bar{u}_m = \sum_{i=1}^m u_i / m. \quad (3)$$

The analyst then can use \bar{q}_m to estimate Q and

$$T_d = (n_2 / n_1 - n_2 / N) \bar{u}_m + b_m / m \quad (4)$$

to estimate the variance of \bar{q}_m . Derivations of these estimates are presented in Section 2.1.2. We note that without subsampling, i.e., $n_2 = n_1$, (4) equals the variance estimate for standard partial synthesis (Reiter 2003). For large n_2 , inferences are based on a t -distribution, $(\bar{q}_m - Q) \sim t_{v_d}(0, T_d)$, with degrees of freedom $v_d = (m-1)(1 + (n_2 / n_1 - n_2 / N) m \bar{u}_m / b_m)^2$.

The inferential methods can be extended to stratified samples in which the NSI uses the same strata for the subsample and original sample. Let N_h be the population size in stratum h , where $h = 1, \dots, H$. For each h , let \bar{q}_{mh} and T_{dh} be the values of (1) and (4) computed using only the records in d_{syn} in stratum h . These estimates are used in inferences for population quantities in stratum h . For inferences about the entire population mean, the point estimate of Q is $\bar{q}_m = \sum_h (N_h / N) \bar{q}_{mh}$, and its estimated variance is $T_d = \sum_h (N_h / N)^2 T_{dh}$. Point and variance estimates for nonlinear functions of means can be derived using Taylor series expansions. We note that NSIs should release the values of n_{2h} / n_{1h} for all strata to enable variance estimation.

2.1.2 Derivation of inferences for the different random subsamples approach

The analyst seeks $f(Q | d_{syn})$, which can be written as

$$f(Q | d_{syn}) = \int f(Q | D_{syn}, d_{syn}) f(D_{syn} | d_{syn}) dD_{syn}. \quad (5)$$

For all derivations in Section 2.1.2, we assume that the analyst's distributions are identical to those used by the NSI for creating D_{syn} . We also assume that the sample sizes are large enough to permit normal approximations for these distributions. Thus, we require only the first two moments for each distribution, which we derive using standard large sample Bayesian arguments. Diffuse priors are assumed for all parameters.

Let Q_i and U_i be the point estimate of Q and its variance that the analyst would compute with D_i (which is not available to the analyst). Let \bar{Q}_m , \bar{U}_m , and B_m be defined as in (1)-(3) but using Q_i and U_i . From standard partial synthesis results (Reiter 2003), we have $(Q | D_{syn}) \sim N(\bar{Q}_m, \bar{U}_m + B_m / m)$. We assume that $(q_i | D_i) \sim N(Q_i, (1 - n_2 / n_1) u_i)$ and, as is typical in multiple imputation contexts, that $u_i \approx \bar{u}_m$. Thus, using standard Bayesian theory, we have $(\bar{Q}_m | d_{syn}) \sim N(\bar{q}_m, (1 - n_2 / n_1) \bar{u}_m / m)$ and $((m-1)b_m / (B_m + (1 - n_2 / n_1) \bar{u}_m) | d_{syn}) \sim \chi_{m-1}^2$. Hence, we have $f(Q | d_{syn}, B_m, \bar{U}_m) = N(\bar{q}_m, \bar{U}_m + B_m / m + (1 - n_2 / n_1) \bar{u}_m / m)$.

To get $f(Q | d_{syn})$, we need to integrate out B_m and \bar{U}_m from this distribution. We do so by substituting B_m and \bar{U}_m with their approximate expected values. To approximate

$E(B_m | d_{syn})$, we use $b_m - (1 - n_2/n_1) \bar{u}_m$. To approximate $E(\bar{U}_m | d_{syn})$, we note that

$$\begin{aligned} \text{Var}(Q | d_i) &= E[\text{var}(Q | D_i) | d_i] + \text{var}[E(Q | D_i) | d_i] \\ &= E(U_i | d_i) + \text{var}(Q_i | d_i). \end{aligned} \quad (6)$$

Here, $\text{var}(Q | d_i) = (1 - n_2/N)u_i$. Solving (6), we have $E(\bar{U}_m | d_{syn}) \approx (n_2/n_1 - n_2/N)\bar{u}_m$. After substitution of these expected values, we have $\text{var}(Q | d_{syn}) = T_d$.

Since we use an estimated variance for Q , we approximate $f(Q | d_{syn})$ with a t -distribution with mean \bar{q}_m and variance T_d . The degrees of freedom, v_d , is derived by matching the first two moments of $(v_d T_d) / \{(n_2/n_1 - n_2/N)\bar{u}_m + B_m/m + (1 - n_2/n_1)\bar{u}_m/m\}$ to those of a $\chi^2_{v_d}$ distribution.

2.2 Releasing the same random subsample

At first glance, releasing a common set of records in each subsample looks like standard partial synthesis. However, Reiter's (2003) variance estimator can be positively biased in this context. To illustrate, suppose that D comprises one variable with sample mean \bar{x}_i . Also suppose that we create D_{syn} by replacing all values of x , and we randomly select a common set of n_2 records for the subsample. Let $m = \infty$, and let Q be the population mean of x . If replacements are simulated from the correct model, which is estimated with D , then $\bar{q}_\infty = \bar{x}_i$. Hence, $\text{var}(\bar{q}_\infty)$ is identical to the variance of \bar{x}_i , which is $(1 - n_1/N)s_1^2/n_1$. However, Reiter's (2003) variance estimate includes \bar{u}_m based on $(1 - n_2/N)s_2^2/n_2$, where $E(s_2^2) = s^2$. Hence, in general Reiter's (2003) variance will have positive bias for subsamples with synthesis.

In place of standard partial synthesis, we adopt the approach taken by Reiter (2008) for multiple imputation for missing data when records used for imputation are not used or disseminated for analysis. This setting is akin to subsampling the same records in each d_i because the models for the synthesis are estimated with D , but the analyst only has d_{syn} for analysis; that is, not all records used for imputation are disseminated for analysis.

For convenience, we summarize the methodology of Reiter (2008) here but do not include the derivations. First, as in standard partial synthesis, the NSI estimates the synthesis models using only the records whose values will be synthesized. Let θ be the parameters that govern the distribution of the synthetic data models. Second, the NSI samples m values of θ from its posterior distribution. Third, for each drawn $\theta^{(l)}$ where $l = 1, \dots, m$, the NSI draws a replacement dataset $D^{(l,p)}$ from the synthesis models based on $\theta^{(l)}$. The NSI repeats this process r times for each $\theta^{(l)}$. Finally, the NSI releases the collection of $M = mr$

subsamples from these datasets, $d^* = \{d^{(l,p)}: l = 1, \dots, m; p = 1, \dots, r\}$. Each $d^{(l,p)}$ includes an index of its nest l .

For $l = 1, \dots, m$ and $p = 1, \dots, r$, let $q^{(l,p)}$ and $u^{(l,p)}$ be the estimate of Q and its estimated variance computed with $d^{(l,p)}$. Here, $u^{(l,p)}$ includes the finite population correction factor. The following quantities are used for inferences:

$$\bar{q}_M = \sum_{l=1}^m \sum_{p=1}^r q^{(l,p)} / (mr) = \sum_{l=1}^m \bar{q}_r^{(l)} / m, \quad (7)$$

$$\bar{w}_M = \sum_{l=1}^m \sum_{p=1}^r (q^{(l,p)} - \bar{q}_r^{(l)})^2 / \{m(r-1)\} = \sum_{l=1}^m w_r^{(l)} / m, \quad (8)$$

$$b_M = \sum_{l=1}^m (\bar{q}_r^{(l)} - \bar{q}_M)^2 / (m-1), \quad (9)$$

$$\bar{u}_M = \sum_{l=1}^m \sum_{p=1}^r u^{(l,p)} / (mr). \quad (10)$$

The analyst can use \bar{q}_M to estimate Q and $T_s = \bar{u}_M - \bar{w}_M + (1+1/m)b_M - \bar{w}_M/r$ to estimate the variance of \bar{q}_M . When r is large, inferences are based on a t -distribution, $(\bar{q}_M - Q) \sim t_{v_s}(0, T_s)$, with degrees of freedom

$$v_s = \left(\frac{\{(1+1/m)b_M\}^2}{(m-1)T_s^2} + \frac{\{(1+1/r)\bar{w}_M\}^2}{\{m(r-1)\}T_s^2} \right)^{-1}. \quad (11)$$

It is possible that $T_s < 0$, particularly for small m and r . Instead, analysts can use the always positive but conservative variance estimator, $T_s^* = \lambda T_s + (1-\lambda)(1+1/m)b_M$, where $\lambda = 1$ when $T_s > 0$ and $\lambda = 0$ otherwise. Motivation for this estimator is provided in Reiter (2008). Generally, negative values of T_s can be avoided by making m and r large. When $T_s < 0$, inferences are based on a t -distribution with $(m-1)$ degrees of freedom, which comes from using only the first term and T_s^* in (11).

For stratified designs, the point estimate for whole population quantities is $\bar{q}_M = \sum_h (N_h/N) \bar{q}_{Mh}$, and its estimated variance is $T_s = \sum_h (N_h/N)^2 T_{sh}$, where \bar{q}_{Mh} and T_{sh} are the point estimate and its variance in stratum h . The degrees of freedom in the t -distribution for stratified sampling is

$$\begin{aligned} v_{st} &= \left\{ \frac{\left\{ \sum_h ((N_h/N)^2 (1+1/m)b_{Mh}) \right\}^2}{(m-1) \sum_h (N_h/N)^2 T_{sh}^2} \right. \\ &\quad \left. + \frac{\left\{ \sum_h ((N_h/N)^2 (1+1/r)\bar{w}_{Mh}) \right\}^2}{\{m(r-1)\} \sum_h (N_h/N)^2 T_{sh}^2} \right\}^{-1}. \end{aligned} \quad (12)$$

This is derived by moment matching to a χ^2 random variable.

3. Illustrative simulations using a stratified sampling design

In this section, we investigate the analytical properties of the inferential procedures for subsampling with synthesis for stratified simple random sampling. We generate a population of $N = 1,000,000$ records comprising five variables, Y_1, \dots, Y_5 , in $H = 4$ strata. Y_1 is a categorical variable with ten categories generated according to the distribution in Table 1. The distributions for (Y_2, \dots, Y_5) are displayed in Table 2, along with the stratum sizes.

Table 1
Empirical distribution of Y_1 in the generated population

	1	2	3	4	5	6	7	8	9	10
percentage	24.77	32.63	16.38	15.06	7.13	2.53	0.95	0.33	0.15	0.09

To create D , we randomly sample $n_{1h} = 7,500$ records from each stratum. Each subsample comprises $n_{2h} = 5,000$ records for each stratum. In practice, the NSI might use proportional allocation to set each n_{1h} and choose smaller sampling rates to set n_{2h} . We use a common sample size and large sampling fractions to illustrate that the variance formulas for subsampling with synthesis correctly handle non-trivial finite population correction factors, e.g., 50% of the records are sampled in stratum 4.

We consider Y_4 and Y_5 to be the confidential variables and illustrate two synthesis scenarios. In the first, we

synthesize all records' values of Y_4 and Y_5 . To do so, in each stratum we simulate Y_{4h} using a regression of Y_{4h} on (Y_{1h}, Y_{2h}, Y_{3h}) estimated with D , and we simulate Y_{5h} using a regression of Y_{5h} on $(Y_{1h}, Y_{2h}, Y_{3h}, Y_{4h})$ estimated with D . Predictions of Y_{5h} are based on the synthesized values of Y_{4h} . In the second approach, in each stratum we replace Y_{4h} and Y_{5h} only for all records with $Y_{3h} > p_h$, where p_h is the 90th percentile of Y_3 in the population in stratum h . We generate replacement values by sampling from regression models; however, the models in each stratum are estimated only with those records satisfying $Y_{3h} > p_h$.

For the different subsamples approach, we generate $m = 5$ synthetic surveys as outlined in Section 2.1. For the same subsample approach, we first draw $m = 5$ values of θ , the regression coefficients and variances. For each $\theta^{(i)}$, we generate $r = 5$ synthetic datasets for every first stage nest.

For all scenarios, we repeat the process of (i) creating D by sampling from the population and (ii) generating subsamples with synthesis a total of 5,000 times. For each of these 5,000 runs, we obtain inferences for fifty quantities, including the population means and within-stratum means of Y_4 and Y_5 , the coefficients from a regression of Y_3 on all other variables, and the coefficients from a regression of Y_5 on all other variables. The regressions are estimated separately in each stratum.

Table 2
Parameters for drawing (Y_2, \dots, Y_5) for the population

	Stratum size	Model	Distribution of the error term
Stratum 1	750,000	$Y_2 = Y_1 + e$ $Y_3 = Y_1 + Y_2 + e$ $Y_4 = Y_1 + Y_2 + Y_3 + e$ $Y_5 = Y_1 + Y_2 + Y_3 + Y_4 + e$	$e \sim N(0, 5)$
Stratum 2	200,000	$Y_2 = 2Y_1 + e$ $Y_3 = 2Y_1 + 0.5Y_2 + e$ $Y_4 = 2Y_1 + 0.5Y_2 + Y_3 + e$ $Y_5 = 2Y_1 + 0.5Y_2 + 0.5Y_3 - 0.25Y_4 + e$	$e \sim N(0, 10)$
Stratum 3	40,000	$Y_2 = -3Y_1 + e$ $Y_3 = -3Y_1 - 1.5Y_2 + e$ $Y_4 = -3Y_1 + Y_2 - 1 / 3Y_3 + e$ $Y_5 = -3Y_1 + Y_2 - 1 / 3Y_3 + 1 / 9Y_4 + e$	$e \sim N(0, 30)$
Stratum 4	10,000	$Y_2 = -2Y_1 + e$ $Y_3 = -Y_1 - 1.5Y_2 + e$ $Y_4 = -2Y_1 + Y_2 + 1 / 4Y_3 + e$ $Y_5 = 2Y_1 - Y_2 - 1 / 4Y_3 + 1 / 16Y_4 + e$	$e \sim N(0, 20)$

Figure 1 displays key results of the simulations. The left panel displays the ratios of the simulated average of T_d (and T_s) over the corresponding simulated $\text{var}(\bar{q}_m)$ for the fifty estimands. The median ratios are close to one in all scenarios, and the averages of T_d (and T_s) never differ by more than 10% from their actual variances. Thus, both T_s and T_d appear to be approximately valid variance estimators.

The middle panel of Figure 1 summarizes the percentages of the 5,000 synthetic 95% confidence intervals based on T_d (and on T_s) that cover their corresponding Q . The coverage rates are close to 0.95 except for the regression coefficients for the same subsampling approach with 100% synthesis. For these coefficients, $T_s < 0$ in up to 38% of the simulation runs, so that confidence intervals are based on the conservative T_s^* . The highest fraction of negative variances occurs in the smallest stratum which has a sampling rate of 50%. All variance estimates are positive when only 10% of the records are synthesized.

The right panel of Figure 1 displays the ratios of the simulated root mean squared error (RMSE) of \bar{q}_m over the simulated RMSE from the subsamples without any synthesis. For the same subsampling approach, the RMSEs of the synthetic subsamples tend to be smaller than the RMSEs based on the subsamples without any synthesis, particularly for the 100% synthesis. The smaller RMSEs result because the synthesis models are determined with D , i.e., the survey data before taking the subsample, so that they carry additional information that is not in the subsamples without

synthesis. For the different synthetic subsamples, the RMSE ratios typically exceed one. Here, increased synthesis leads to greater loss in efficiency. We note that the RMSEs from the different sample and same sample approaches in Figure 1 are not directly comparable because they are based on different denominators.

To enable comparisons across the methods, as well as to illustrate the losses in efficiency from subsampling, we repeat the simulation design using $m = 25$ for the independent subsamples approach and $mr = 25$ for the same subsamples approach. The left panel of Figure 2 displays the simulated RMSE ratios for the fifty estimands in the different scenarios, where the denominators are the average RMSEs based on the original data before any confidentiality protection. The right panel of Figure 2 displays the ratios of simulated average lengths of the 95% confidence intervals, where the denominators are the average lengths based on the original data before any confidentiality protection. Based on the left panel, for a given total number of released datasets and given synthesis percentage, the independent sample approach results in more efficient estimates than the same sample approach. The right panel tells a similar story, although it is harder to see because of the scaling. Here, the same sample approach with 100% synthesis results in high fractions of negative variance estimates, so that the adjusted variance T_s^* is often used, thereby inflating the interval lengths. Figure 2 also includes results from synthesis without any subsampling, which generally provides more efficient estimates than either subsampling approach.

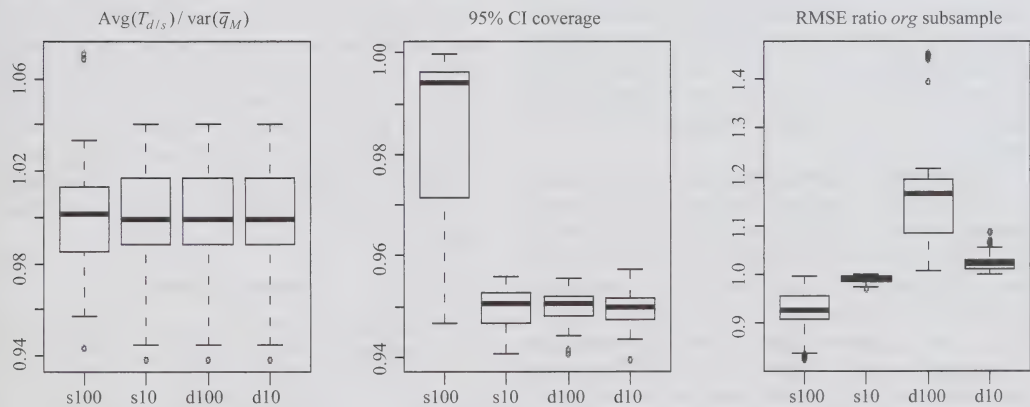


Figure 1 Simulation results for the stratified sampling design. In the labels, s and d indicate the same subsample and the different subsamples approach. The numbers indicate the percentage of records that are being synthesized. The denominators of the RMSE are based on the point estimates from the subsamples without synthesis. For the different subsamples approach, the RMSE is computed from the average of the m point estimates. Each box plot comprises fifty estimands

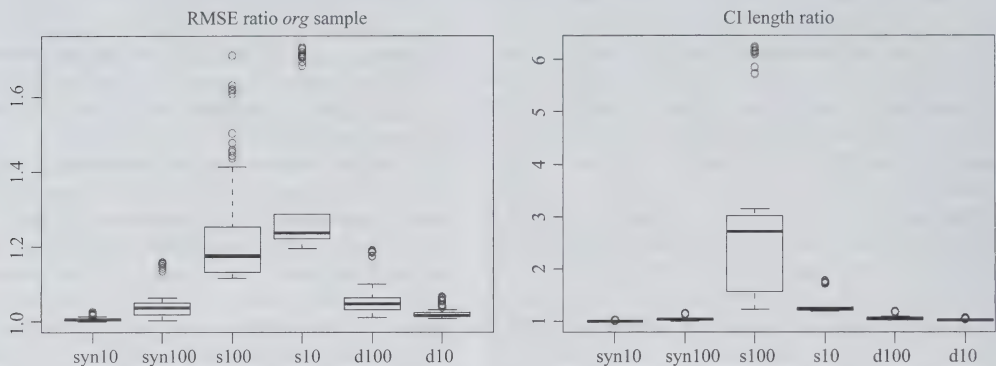


Figure 2 Efficiency comparisons for the stratified sampling design. In the labels, *org* and *syn* indicate the original sample and the synthetic sample before subsampling; and, s, d, and the numbers are as in Figure 1. The denominators of the RMSE are based on the point estimates from the original sample without synthesis. Each box plot comprises fifty estimands

4. Concluding remarks

The different subsamples and same subsamples approaches have competing advantages. For a fixed number of released datasets M , the different subsamples approach enables estimation with greater efficiency than the same subsamples approach - as evident in Figure 2 - since the released subsamples are independent rather than correlated. The different subsamples approach also guarantees positive variance estimates; the same subsample approach does not. However, with large M the different subsamples approach weakens the confidentiality protections of subsampling, since the combined datasets are likely to contain most of the records from the original survey. Hence, unless the subsampling rate is small (e.g., 1% or 2%), the NSI may have to make m modest (e.g., $m = 5$) to use the different subsamples approach. Because of this, the different samples approach is not viable when the original sample size is modest.

As an alternative to subsampling with synthesis, agencies could release partially synthetic data that include all records from the original sample, assuming that they are willing to release files of that size. Partial synthesis on the original data generally engenders estimates with lower variances than subsampling with synthesis - as evident in Figure 2 - since more records are released. However, partial synthesis on the original data generally engenders higher disclosure risks than subsampling with synthesis, since more at risk records are in the released data and since the additional protection from subsampling is absent. Agencies can compare the two options on disclosure risks using the methods of Drechsler and Reiter (2008), which account for the protection afforded

by sampling, and on data utility by comparing inferences for representative analyses.

It is also possible that the process of subsampling may engender sufficient additional protection to enable lesser amounts of synthesis than would be necessary in a partial synthesis of the entire original dataset. Evaluating the data utility for subsampling with synthesis versus synthesis only for given disclosure risks is beyond the scope of this short note, but it is an interesting area for future research.

We have not developed subsampling with synthesis approaches for sampling designs other than (stratified) simple random samples. For the different subsamples approach, appropriate inferential methods require an approximately unbiased estimate of the variance from the first phase of sampling that can be computed from the subsample alone. This is elusive for complicated designs. For the same subsample approach, we conjecture that analysts can use the inferential methods presented in Section 2.2, provided that \bar{u}_M appropriately accounts for the two phases of sampling. We note that the formulas for \bar{w}_M and b_M remain the same for other designs. Evaluating this conjecture is a subject of future research.

Acknowledgements

This research was supported by U.S. National Science Foundation grant SES-0751671.

References

- Drechsler, J., and Reiter, J.P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases*, (Eds., J. Domingo-Ferrer and Y. Saygin), New York: Springer, 227-238.

- Drechsler, J., and Reiter, J.P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *Journal of Official Statistics*, 25, 589-603.
- Drechsler, J., and Reiter, J.P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105, 1347-1357.
- Elliott, M., and Purdam, K. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records. *Environment and Planning A*, 39, 1101-1118.
- Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.
- Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181-189.
- Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 235-242.
- Reiter, J.P. (2005). Estimating identification risks in microdata. *Journal of the American Statistical Association*, 100, 1103-1113.
- Reiter, J.P. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*, 95, 933-946.
- Willenborg, L., and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.
- Winkler, W.E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Tech. rep., Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

A hierarchical Bayesian nonresponse model for two-way categorical data from small areas with uncertainty about ignorability

Balagobin Nandram and Myron Katzoff¹

Abstract

We study the problem of nonignorable nonresponse in a two dimensional contingency table which can be constructed for each of several small areas when there is both item and unit nonresponse. In general, the provision for both types of nonresponse with small areas introduces significant additional complexity in the estimation of model parameters. For this paper, we conceptualize the full data array for each area to consist of a table for complete data and three supplemental tables for missing row data, missing column data, and missing row and column data. For nonignorable nonresponse, the total cell probabilities are allowed to vary by area, cell and these three types of "missingness". The underlying cell probabilities (*i.e.*, those which would apply if full classification were always possible) for each area are generated from a common distribution and their similarity across the areas is parametrically quantified. Our approach is an extension of the selection approach for nonignorable nonresponse investigated by Nandram and Choi (2002a, b) for binary data; this extension creates additional complexity because of the multivariate nature of the data coupled with the small area structure. As in that earlier work, the extension is an expansion model centered on an ignorable nonresponse model so that the total cell probability is dependent upon which of the categories is the response. Our investigation employs hierarchical Bayesian models and Markov chain Monte Carlo methods for posterior inference. The models and methods are illustrated with data from the third National Health and Nutrition Examination Survey.

Key Words: Metropolis-Hastings sampler; SIR algorithm; Nonignorable nonresponse model; Expansion model.

1. Introduction

In sample surveys, data are typically summarized in two-way categorical tables. We consider the problem of nonignorable nonresponse for many $r \times c$ categorical tables, each obtained from a single area. For many of these surveys, there are missing data and this gives rise to partial classification of the sampled individuals. Thus, for each two-way table there are both item nonresponse (one of the two categories is missing) and unit nonresponse (both categories are missing). One may not know how the data are missing and a model that includes some difference between the observed data and missing data (*i.e.*, nonignorable missing data) may be preferred. For a general $r \times c$ categorical table, we address the issue of estimation of the cell probabilities of the two-way tables when there is possibly nonignorable nonresponse but there is really no information about ignorability. In such a situation, we would like to express a degree of uncertainty about ignorability. Nandram and Choi (2002a, b) have described an expansion model appropriate for binary data when there are data from many small areas. We will extend this work to $r \times c$ categorical tables.

Letting x denote the covariates and y the response variable, Little and Rubin (2002) describe three types of missing-data mechanism. These types differ according to whether the probability of response (a) is independent of x

and y ; (b) depends on x but not on y ; or (c) depends on y and possibly x . The missing data are missing completely at random (MCAR) in (a), missing at random (MAR) in (b) and the data are missing not at random (MNAR) in (c). Models for MCAR and MAR missing-data mechanisms are called ignorable if the parameters of the dependent variable and the response variable are distinct (Rubin 1976). Models for MNAR missing-data mechanisms are called nonignorable. The general difficulty with nonignorable nonresponse model is that the parameters are not identifiable [*e.g.*, see Nandram and Choi (2004, 2005, 2008, 2010) and Nandram, Han and Choi (2002)].

For a $r \times c$ categorical table, let $I_{ijkl} = 1$ if the l^{th} individual within the i^{th} area falls in the j^{th} row and k^{th} column and 0 otherwise. Also, let $J_{il} = 1$ if the l^{th} individual within the i^{th} area has complete information and 0 otherwise. Finally, let $P(J_{il} = 1 \mid I_{ijkl} = 1, I_{ijk'l} = 0, j' \neq j, k' \neq k) = \pi_{ijk}$. For unit nonresponse, if $\pi_{ijk} = \pi_i$, the model is ignorable; for item nonresponse, if the columns are missing, row is observed and $\pi_{ijk} = \pi_j$ (or $\pi_{ijk} = \pi_i$), the model is ignorable; and if the rows are missing but columns are observed and $\pi_{ijk} = \pi_k$ (or $\pi_{ijk} = \pi_i$), the model is ignorable. All other models are nonignorable; see Rubin (1976) for further explanation.

Nandram and Choi (2002a, b) use an expansion model to study nonignorable nonresponse binary data. The expansion

1. Balagobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609. E-mail: balnan@wpi.edu; Myron Katzoff, Office of research and Methodology, National Center for Health Statistics, CDC, 3311 Toledo Road, Hyattsville, MD 20782. E-mail: mjks5@cdc.gov.

model, a nonignorable nonresponse model, degenerates into an ignorable nonresponse model (in the spirit of Draper 1995) when a centering parameter is set to unity. This permits an expression of uncertainty about ignorability; see also Forster and Smith (1998).

We discuss the model of Nandram and Choi (2002a, b) for binary data from small areas. So that J_{il} denote the response indicators and I_{il} denote the binary response. Specifically, introducing the centering parameters γ_i for area i to incorporate uncertainty about ignorability, the model of Nandram and Choi (2002a, b) is

$$\begin{aligned} I_{il} | p_i &\stackrel{iid}{\sim} \text{Bernoulli}(p_i), \\ J_{il} | \{\pi_i, J_{il} = 0\} &\stackrel{iid}{\sim} \text{Bernoulli}(\pi_i), l = 1, \dots, n_i, i = 1, \dots, I, \\ J_{il} | \{\pi_i, \gamma_i, y_{il} = 1\} &\stackrel{iid}{\sim} \text{Bernoulli}(\gamma_i \pi_i), 0 < \gamma_i \pi_i < 1. \end{aligned}$$

When $\gamma_i = 1$, the nonignorable nonresponse model degenerates to an ignorable nonresponse model. Here γ_i is the ratio of the odds of success among respondents to the odds of success among all individuals for the i^{th} area. The parameter γ_i describes the extent of nonignorability of the response mechanism for area i , and it is through the γ_i that uncertainty about ignorability is incorporated. Nandram and Choi (2002a, b) define $\delta_i = \pi_i \{\gamma_i p_i + (1 - p_i)\}$ to be the probability that an individual responds in area i in the entire population, and with a belief that all the areas are similar they take $(p_i, \delta_i, \gamma_i)$ to have a common distribution. A priori they take beta distributions for p_i and π_i respectively.

Here, the parameters are not identifiable. However, if $\gamma_i = 1$, then all the parameters are identifiable. That is, identifiability of the parameters depend on the γ_i . Note, that when $\gamma_i = 1$, we get an ignorable model for a MAR mechanism. As the parameters are identifiable in this model, it is quite sensible to use this model (or similar models) as a baseline model. However, note this model is still not justified because it assumes that missing data are like observed data. Thus, to add flexibility to this ignorable nonresponse model, we use the γ_i .

Let γ_{iuv} be the number of individuals with $I_{il} = u$, $J_{il} = v$ ($u, v = 0, 1$) in the i^{th} area. Then, under the model,

$$(y_{i00}, y_{i01}, y_{i10}, y_{i11}) | \pi_i, p_i, \gamma_i \stackrel{\text{ind}}{\sim} \text{Multinomial}\{n_i, (1 - p_i)(1 - \pi_i), (1 - p_i)\pi_i, (1 - \gamma_i \pi_i)p_i, \gamma_i \pi_i p_i\}$$

with independence over areas. Here, only y_{i01} and y_{i11} are observed, and therefore all parameters are nonidentifiable if the γ_i are unknown. We obtain the likelihood function in a similar manner for the more complete $r \times c$ categorical table with missing data.

We start with a gamma distribution, and to permit centering on the ignorable nonresponse model, we must take each γ_i to have mean 1. However, we need to use a truncated gamma distribution because $0 < \pi_i < 1$ and $0 < \gamma_i \leq 1/\pi_i$. An interesting idea of Nandram and Choi (2002a, b) is to model the centering as a truncated gamma

$$\gamma_i | v \stackrel{iid}{\sim} \text{Gamma}(v, v), 0 < \gamma_i < 1/\pi_i, 0 < \pi_i < 1.$$

The model is complete with noninformative prior densities on all hyperparameters. One can use alternative distributions (e.g., a truncated lognormal density) for the γ_i , but this is not a key issue and it would not matter much.

One can use an area level model with random effects in which, conditional on the observed data, the nonresponse is dependent upon area-level random effects. This can be formulated using a logit link function, but we have not developed our models in this direction partly because we are not using covariates here; see Nandram and Choi (2010) for the use of covariates and random effects.

The approach in Nandram and Choi (2002a, b) is attractive, but it does not apply immediately to the current $r \times c$ categorical table problem. Specifically, only one centering parameter per area is needed in Nandram and Choi (2002a, b). In our formulation, one now needs rc centering parameters per area; each of these parameters has to have a distribution centered at one to allow degeneration to the ignorable nonresponse model. There are also inequality constraints that must be included in the non-ignorable nonresponse model. In addition, one cannot rule out the possibility that these parameters are correlated. The methodology needed to apply the work of Nandram and Choi (2002a, b) to the $r \times c$ categorical table is not straightforward. Noting these difficulties Nandram, Liu, Choi and Cox (2005) (with a single supplemental table) and Nandram, Cox and Choi (2005) (with the three supplemental tables) use a simpler idea, but not quite as elegant as in Nandram and Choi (2002a, b), for centering; see also Nandram and Choi (2005).

Essentially, Nandram, Cox and Choi (2005) and Nandram, Liu, Cox and Choi (2005) assume an ignorable model, obtain samples of the response probabilities and use these sampled response probabilities to fit the response probabilities of a nonignorable nonresponse model while "controlling" its parameters. Of course, a possible alternative occurs when there is information about the degree of nonignorability. However, the problem of incorporating prior information about a systematic departure from ignorability is more complex for our problem, and it would need additional costly field work to obtain such information.

We discuss our philosophy about the nonignorable nonresponse problem, a fundamentally aliased problem. In fact, this problem is extremely difficult and we believe that

there is really no solution to it, but we must try. Without any information one cannot tell how respondents and non-respondents differ. An ignorable nonresponse model is short because it assumes that respondents and nonrespondents are similar, but the respondents and nonrespondents may differ. Statisticians must not confront imprecision (sampling error) only, but they must be bold enough to study subjectivity (ignorance arising from missing information). Unfortunately, as is well known, nonignorable nonresponse models have nonidentifiable parameters. We discuss how the key nonignorability parameters are identified. We know that if the respondents and nonrespondents are similar, then the γ_i are equal unity, and we get the ignorable nonresponse model with all parameters identified. We can now expand the ignorable nonresponse model into a nonignorable nonresponse model by putting a distribution on these γ_i centered at 1, still maintaining identifiability. One can formulate a nonignorable nonresponse model to add flexibility to the ignorable nonresponse model as we have done in our work; the flexibility is a form of sensitivity analysis, coherent in this case, and indeed it is a Bayesian uncertainty (risk) assessment (e.g., Greenland 2009). This is what we have been doing or trying to do in our work.

In this paper we attempt to solve the difficult problem of Nandram and Choi (2002a, b) in its original form for $r \times c$ tables for many areas. The plan of this paper is as follows. In Section 2, we describe the hierarchical Bayesian model. Specifically we describe the nonignorable nonresponse mechanism and we construct an appropriate prior distribution. In Section 3, we show how to fit the model using the sampling importance resampling (SIR) algorithm to subsample from an approximate posterior density after an innovative collapsing of the complete joint posterior density. In Section 4, we illustrate our methodology with public-use data from thirteen states in the third National Health and Nutrition Examination Survey (NHANES III). Section 5 has concluding remarks.

2. The nonignorable nonresponse model

For the problem of nonresponse in a two-dimensional table, we can have both item and unit nonresponse. Thus, one may consider the full data array to consist of four tables: one for complete data and three supplemental tables - one for missing row information, one for missing column information and a table for which neither row nor column membership has been recorded. Throughout this paper, we index rows by $j = 1, \dots, r$; columns, by $k = 1, \dots, c$; and the four tables by $s = 1, 2, 3, 4$. We index areas by $i = 1, 2, \dots, A$ and individuals within areas by $l = 1, 2, \dots, n_i$. We next describe the nonignorable nonresponse model (i.e., the expansion model).

2.1 Sampling process

We adapt the terminology and definitions used in Nandram, Cox and Choi (2005) to our situation. For sample individual l in area i , let

$$I_{ijkl} = \begin{cases} 1, & \text{if the outcome category is } (j, k) \\ 0, & \text{otherwise,} \end{cases}$$

and let \mathbf{J}_{il} denote one of the 4-tuples $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, $(0, 0, 0, 1)$. We assume that

$$\begin{aligned} \mathbf{I}_{il} &\stackrel{\text{def}}{=} \text{vec}(\{I_{ijkl} | j = 1, \dots, r; \\ &\quad k = 1, \dots, c\}) | \mathbf{p}_i \sim \text{Mult}\{1, \mathbf{p}_i\} \end{aligned} \quad (1)$$

and

$$\begin{aligned} \mathbf{J}_{il} &| \{I_{ijkl} = 1, I_{ij'k'l} = 0 \text{ for all } j' \neq j \\ &\text{and } k' \neq k | \pi_{ijk}\} \stackrel{\text{iid}}{\sim} \text{Mult}\{1, \pi_{ijk}\}, \end{aligned} \quad (2)$$

where $\mathbf{p}_i \stackrel{\text{def}}{=} \text{vec}(\{p_{ijk} | j = 1, 2, \dots, r; k = 1, 2, \dots, c\})$ is a vector of probabilities for the table of rc categories for the variable of observation which must sum to one and, for cell (j, k) in that two-dimensional table,

$$\pi_{ijk} \stackrel{\text{def}}{=} \text{vec}(\{\pi_{isjk}\} \text{ for } s = 1, 2, 3, 4)$$

is a vector of probabilities which must sum to one.

Next, we define cell counts y_{isjk} , for each table $s = 1, \dots, 4$ for area i such that, for cell (j, k) ,

$$(y_{i1jk}, y_{i2jk}, y_{i3jk}, y_{i4jk}) = \sum_{l=1}^{n_i} I_{ijkl} \mathbf{J}_{il},$$

where y_{i1jk} are observed and y_{isjk} , for $s = 2, 3, 4$, are latent variables which satisfy the observed constraints $\sum_k y_{i2jk} = u_{ij}$, $\sum_j y_{i3jk} = v_{ik}$ and $\sum_{j,k} y_{i4jk} = w_i$. All inferences will be conditional on the observed quantities, u_{ij} , v_{ik} and w_i . But see Nandram (2009) for the analysis of a single $r \times c$ table under nonresponse when the margins are also random. We will denote the vector of the y_{i1jk} by \mathbf{y}_1 , the vector of the y_{isjk} , $s = 2, 3, 4$, by $\mathbf{y}_{(1)}$, and the complete vector by $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_{(1)})'$.

The parameters π_{isjk} are not identifiable. If the distributions of these parameters are known completely, then the nonidentifiability will disappear. Thus, the key issue is how to identify these parameters. We know that if the respondents and nonrespondents are similar (i.e., the four patterns, complete and partially complete tables), then we can take $\pi_{isjk} = \pi_{is}$; this is the ignorable nonresponse model. The π_{is} can be estimated by the proportions of cases falling in the four tables for each area. This is a natural point

to start. To expand the ignorable nonresponse model into a nonignorable model, and still maintain identifiability, first we need a simplification. We take $\pi_{ijks} = \psi_{ijk} \pi_{is}$, which gives a nonignorable nonresponse model in which the parameters ψ_{ijk} are not identifiable.

To center the nonignorable model on the ignorable model, we take

$$\pi_{isjk} = \begin{cases} \tilde{\psi}_{ijk} \pi_{is}, & \text{for } s=1, \\ \psi_{ijk} \pi_{is}, & \text{for } s=2,3,4, \end{cases} \quad (3)$$

and require that $\sum_{s=1}^4 \pi_{is} = 1$. A little algebra then yields the relationship

$$\begin{aligned} \tilde{\psi}_{ijk} \pi_{i1} &= \left[1 + (1 - \psi_{ijk}) \left(\frac{1 - \pi_{i1}}{\pi_{i1}} \right) \right] \pi_{i1} \\ &= a_{ijk}(\pi_{i1}, \psi_{ijk}) \psi_{ijk} \pi_{i1}, \end{aligned} \quad (4)$$

where $a_{ijk}(\pi_{i1}, \psi_{ijk}) = \{\psi_{ijk}^{-1} + (\psi_{ijk}^{-1} - 1)(\pi_{i1}^{-1} - 1)\}$, from which it is clear that $\tilde{\psi}_{ijk} = 1$ if, and only if, $\psi_{ijk} = 1$. Note that since $0 \leq \pi_{isjk} \leq 1$ and $(1 - \pi_{i1})^{-1} \leq \min\{\pi_{is}^{-1}; s=2,3,4\}$, it follows that $0 < \psi_{ijk} \leq (1 - \pi_{i1})^{-1}$.

By combining (1) and (2) and noting the definition of π_{isjk} in (3), similar to binary case, we get a multinomial distribution for \mathbf{y} conditional on $\boldsymbol{\pi}, \boldsymbol{\psi}, \mathbf{p}$, and the likelihood function for the sample can now be seen to be

$$\begin{aligned} f(\mathbf{y} | \boldsymbol{\pi}, \boldsymbol{\psi}, \mathbf{p}) &= \prod_{i=1}^A \binom{n_i}{\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \mathbf{y}'_{i3}, \mathbf{y}'_{i4}} \left\{ \left[\prod_{j,k} (\tilde{\psi}_{ijk} \pi_{i1} p_{ijk})^{y_{ijk}} \right] \right. \\ &\quad \left. \prod_{s=2}^4 \prod_{j,k} (\psi_{ijk} \pi_{is} p_{ijk})^{y_{ijk}} \right\} \\ &= \prod_{i=1}^A \binom{n_i}{\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \mathbf{y}'_{i3}, \mathbf{y}'_{i4}} \left\{ \prod_{s=1}^4 \prod_{j,k} (\psi_{ijk} \pi_{is} p_{ijk})^{y_{ijk}} \right. \\ &\quad \left. \prod_{j,k} [a_{ijk}(\pi_{i1}, \psi_{ijk})]^{y_{ijk}} \right\}, \end{aligned} \quad (5)$$

where

$$\begin{aligned} \mathbf{y}_{is}^{rc \times 1} &= \text{vec}(\{y_{isjk} | j=1, \dots, r; k=1, \dots, c\}), \\ \mathbf{y} &= (\mathbf{y}'_{11}, \mathbf{y}'_{12}, \mathbf{y}'_{13}, \mathbf{y}'_{14}, \mathbf{y}'_{21}, \dots, \mathbf{y}'_{24}, \dots, \mathbf{y}'_{A1}, \mathbf{y}'_{A2}, \mathbf{y}'_{A3}, \mathbf{y}'_{A4})', \\ \boldsymbol{\pi}^{4 \times 1} &= (\pi_{11}, \dots, \pi_{14}, \pi_{21}, \dots, \pi_{24}, \dots, \pi_{A1}, \dots, \pi_{A4})', \\ \boldsymbol{\psi}^{A \times c \times 1} &= (\psi_{111}, \dots, \psi_{1rc}, \psi_{211}, \dots, \psi_{2rc}, \dots, \psi_{A11}, \dots, \psi_{Arc}), \\ \text{and} \\ \mathbf{p}^{A \times c \times 1} &= (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_A)'. \end{aligned}$$

Collecting factors which are powers of π_{is} , the likelihood function may also be expressed as

$$\begin{aligned} f(\mathbf{y} | \boldsymbol{\pi}, \boldsymbol{\psi}, \mathbf{p}) &= \prod_{i=1}^A \binom{n_i}{\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \mathbf{y}'_{i3}, \mathbf{y}'_{i4}} \\ &\quad \left\{ \prod_{s=1}^4 \pi_{is}^{y_{is}} \times \prod_{j,k} \{p_{ijk} \psi_{ijk}\}^{y_{ijk}} [a_{ijk}(\pi_{i1}, \psi_{ijk})]^{y_{ijk}} \right\}, \end{aligned} \quad (6)$$

where $0 \leq \pi_{is} \leq 1, \sum_s \pi_{is} = 1$ and $0 \leq \psi_{ijk} \leq (1 - \pi_{i1})^{-1}$. Here we note that y_{is} and y_{ijk} are observed variables but the $y_{i,jk}$ are latent variables.

2.2 Prior construction

The following assumptions describe the prior distributions for the nonignorable nonresponse model:

1. For the vector of cell probabilities \mathbf{p}_i , we assume that

$$\mathbf{p}_i | \boldsymbol{\mu}_i, \tau_i \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_i \tau_i),$$

where $\boldsymbol{\mu}_i = (\mu_{i11}, \mu_{i12}, \dots, \mu_{i1rk}, \mu_{i21}, \dots, \mu_{i1rc})'$; $\mu_{ijk} \geq 0$; and $\sum_{j=1}^r \sum_{k=1}^c \mu_{ijk} = 1$. The parameter τ_i informs us of similarity among the \mathbf{p}_i : the larger τ_i , the more alike the \mathbf{p}_i . This is true because large τ_i means that the variances of the \mathbf{p}_i are small, and because they have the same mean, this means that they are more similar with larger τ_i .

Thus, the density for \mathbf{p} is

$$\begin{aligned} g_1(\mathbf{p} | \boldsymbol{\mu}_i, \tau_i) &= \prod_{i=1}^A g_{1i}(\mathbf{p}_i | \boldsymbol{\mu}_i, \tau_i) \\ &= \prod_{i=1}^A \left\{ \frac{\prod_{j,k} p_{ijk}^{\mu_{ijk} \tau_i - 1}}{D(\boldsymbol{\mu}_i \tau_i)} \right\}, \end{aligned} \quad (7)$$

where, for a k -tuple \mathbf{c} and a scalar t

$$D(\mathbf{c}t) = \frac{\prod_{j=1}^k \Gamma(c_j t)}{\Gamma(t)}$$

for $c_j > 0$ and $\sum_{j=1}^k c_j = 1$.

2. Independently of the \mathbf{p}_i , the $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})'$ follow the specification

$$\boldsymbol{\pi}_i \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2 \tau_2),$$

with $\pi_{is} \geq 0$ and $\sum_s \pi_{is} = 1$, where $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22}, \mu_{23}, \mu_{24})'$, $\mu_{2s} \geq 0$, $\sum_{s=1}^4 \mu_{2s} = 1$ and τ_2 is a measure of similarity among the $\boldsymbol{\pi}_i$. Thus, the density for $\boldsymbol{\pi}_i$ is

$$g_{2i}(\boldsymbol{\pi}_i | \boldsymbol{\mu}_2, \tau_2) = \frac{\prod_{s=1}^4 \pi_{is}^{\mu_{2s} \tau_2 - 1}}{D(\boldsymbol{\mu}_2 \tau_2)}. \quad (8)$$

3. For each i , let $\boldsymbol{\psi}_i = (\psi_{i11}, \dots, \psi_{i1c}, \psi_{i21}, \dots, \psi_{i2c}, \dots, \psi_{irc})'$ so that $\boldsymbol{\psi} = (\boldsymbol{\psi}'_1, \dots, \boldsymbol{\psi}'_A)'$. We assume for each i that the ψ_{ijk} are independently and identically

distributed in accordance with a distribution derived from the $\text{Gamma}(\beta, \beta)$, where the support is confined to the open interval $(0, (1 - \pi_{i1})^{-1})$; in other words, the ordinary gamma distribution is truncated as

$$\psi_{ijk} | \beta, \pi_i \sim \text{Gamma}(\beta, \beta) \quad \text{ind.} \\ \text{such that } 0 < \psi_{ijk} < (1 - \pi_{i1})^{-1}.$$

It is worth noting that these ψ_{ijk} are identically distributed over j and k . Again, one can use other distributions such as a truncated lognormal density, but this will make little difference. In this formulation, there is some information about β because the small areas are assumed to share a common effect.

Thus, for area i , the density for ψ_i is

$$g_{3i}(\psi_i | \beta, \pi_i) \\ = \prod_{j=1}^r \prod_{k=1}^c \left\{ \frac{\beta^\beta \psi_{ijk}^{\beta-1} e^{-\beta \psi_{ijk}}}{\Gamma(\beta)} \right\} \bigg/ \int_0^{(1-\pi_{i1})^{-1}} \frac{\beta^\beta \psi_{ijk}^{\beta-1} e^{-\beta \psi_{ijk}}}{\Gamma(\beta)} d\psi_{ijk},$$

For $0 < \psi_{ijk} < (1 - \pi_{i1})^{-1}$. Making the transformation $t_{ijk} = \beta \psi_{ijk}$, one can see that the normalizing constant in the denominator of each of the factors in $g_{3i}(\psi_i | \beta, \pi_i)$ is $G_\beta[\beta(1 - \pi_{i1})^{-1}]$, where $G_\beta(\cdot)$ is the gamma function with scale parameter β . To eliminate the dependence of the range of integration on π_{i1} , let $\phi_{ijk} = (1 - \pi_{i1}) \psi_{ijk}$ and let $\phi_i = (\phi_{i11}, \dots, \phi_{i1c}, \phi_{i21}, \dots, \phi_{i2c}, \dots, \phi_{irc})'$. Then

$$g_{3i}(\phi_i | \beta, \pi_i) \\ = \prod_{j=1}^r \prod_{k=1}^c \left\{ \frac{\beta^\beta}{(1 - \pi_{i1})^\beta} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta \phi_{ijk}}{1 - \pi_{i1}}}}{\Gamma(\beta) G_\beta[\beta(1 - \pi_{i1})^{-1}]} \right\}, \quad (9)$$

for $0 < \phi_{ijk} < 1$. The joint prior for π_i and ϕ_i is just the product of $g_{3i}(\phi_i | \beta, \pi_i)$ and $g_{2i}(\pi_i | \mu_2, \tau_2)$. Thus, the joint prior for $\phi = (\phi_1, \dots, \phi_A)'$ and π is

$$g^*(\pi, \phi | \mu_2, \tau_2, \beta) \stackrel{\text{def}}{=} \prod_{i=1}^A \{g_{3i}(\phi_i | \beta, \pi_i) \cdot g_{2i}(\pi_i | \mu_2, \tau_2)\}.$$

That is

$$g^*(\pi, \phi | \mu_2, \tau_2, \beta) \\ = \prod_{i=1}^A \left\{ \frac{\prod_{s=1}^4 \pi_{is}^{\mu_{2s}-1}}{D(\mu_2 \tau_2)} \prod_{j=1}^r \prod_{k=1}^c \left\{ \frac{\beta^\beta}{(1 - \pi_{i1})^\beta} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta \phi_{ijk}}{1 - \pi_{i1}}}}{\Gamma(\beta) G_\beta[\beta(1 - \pi_{i1})^{-1}]} \right\} \right\}. \quad (10)$$

The description of the model is completed by specifying the assumptions on the hyperparameters. As there are no

conjugate priors, we use shrinkage priors for τ_1, τ_2 and β because these are proper and noninformative. Priors of the form $p(\tau_1) \propto 1/\tau_1$, and specifically proper diffused gamma priors, are discouraged; see, for example, Gelman (2006). Other alternatives are half Cauchy densities and gamma densities (one would need to specify the hyperparameters). Thus, we take

1. τ_1, τ_2 and β have independent shrinkage priors of the form

$$f(x) = \frac{a_0}{(a_0 + x)^2}, \text{ for } x \geq 0,$$

where a_0 is specified; it is standard practice to take $a_0 = 1$.

2. We also assume that $\mu_1 \sim \text{Dirichlet}(1, 1, \dots, 1)$ and $\mu_2 \sim \text{Dirichlet}(1, 1, 1, 1)$.

Let $\Omega = (\beta, \mu_1, \tau_1, \mu_2, \tau_2)$. Then the density for Ω is

$$p(\Omega) = \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2} (rc - 1)! 3!$$

for τ_1, τ_2 and $\beta \geq 0$, $\sum_{j,k} \mu_{1jk} = 1$ and $\sum_{s=1}^4 \mu_{2s} = 1$.

By Bayes' theorem, the joint posterior density is

$$h(\Omega, p, \pi, \phi, y_{(1)} | y_1, u, v, w) \propto \\ f(y | \pi, \phi, p) g_1(p | \mu_1, \tau_1) g^*(\pi, \phi | \mu_2, \tau_2, \beta) p(\Omega) \\ = \prod_{i=1}^A \left\{ \binom{n_i}{y'_{i1}, y'_{i2}, y'_{i3}, y'_{i4}} (1 - \pi_{i1})^{-y_i} \right. \\ \left. \left\{ \prod_{s=1}^4 \pi_{is}^{y_{is}} \times \prod_{j,k} \{p_{ijk} \phi_{ijk}\}^{y_{i,jk}} [a_{ijk}(\pi_{i1}, \phi_{ijk})]^{y_{i,jk}} \right\} \right. \\ \left. \times \left\{ \frac{\prod_{j,k} p_{ijk}^{\mu_{1jk} \tau_1 - 1}}{D(\mu_1 \tau_1)} \right\} \left\{ \frac{\prod_{s=1}^4 \pi_{is}^{\mu_{2s} \tau_2 - 1}}{D(\mu_2 \tau_2)} \right\} \right. \\ \left. \left. \prod_{j=1}^r \prod_{k=1}^c \frac{\beta^\beta}{(1 - \pi_{i1})^\beta} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta \phi_{ijk}}{1 - \pi_{i1}}}}{\Gamma(\beta) G_\beta[\beta(1 - \pi_{i1})^{-1}]} \right\} \right\} \\ \times \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2}, \quad (11)$$

where, substituting $(1 - \pi_{i1})^{-1} \phi_{ijk}$ for ψ_{ijk} ,

$$a_{ijk}(\pi_{i1}, \phi_{ijk}) = \left(\frac{1 - \pi_{i1}}{\phi_{ijk}} \right) \left[1 + \frac{1}{\pi_{i1}} \{1 - \pi_{i1} - \phi_{ijk}\} \right]. \quad (12)$$

To make inferences about the p_{ijk} , we will draw samples from $h(\Omega, p, \pi, \phi, y_{(1)} | y_1, u, v, w)$ using Markov chain Monte Carlo methods. This procedure is described in Section 3.

3. Computations

We use the SIR algorithm to subsample a random sample from an approximate posterior density. There are three steps to accomplish this task. We collapse over the \mathbf{p}_i , $\boldsymbol{\pi}_i$ and $\boldsymbol{\phi}_i$, approximate the collapsed density by a simpler one and sample from it, and then subsample these samples to get samples from the original density. We show how to do these three steps in this section.

To obtain the approximation and to simplify the computations, in Appendix A we collapse over the \mathbf{p}_i , $\boldsymbol{\pi}_i$ and $\boldsymbol{\phi}_i$ to get

$$h(\Omega, \mathbf{y}_{(1)} | \mathbf{y}_1, \mathbf{u}, \mathbf{v}, \mathbf{w}) = \pi_a(\Omega | \mathbf{y}_{(1)} | \mathbf{y}_1, \mathbf{u}, \mathbf{v}, \mathbf{w}) \cdot \prod_{i=1}^A I_i,$$

where

$$I_i = \iiint_0^{1-\pi_{i1}} \frac{G_{rc\beta} \left(\frac{\beta b_i}{1-\pi_{i1}} \right)}{\left[G_{\beta} \left(\frac{\beta}{1-\pi_{i1}} \right) \right]^{rc}} \prod_{j,k} \left\{ \left(\frac{W_i}{\beta} \right) \sum_{s=2}^4 y_{isjk} \left[\frac{1}{\phi_{ijk}^*} \left(1 + \frac{1-\pi_{i1}}{\pi_{i1}} \left\{ 1 - \frac{W_i \phi_{ijk}^*}{\beta} \right\} \right) \right]^{y_{i1jk}} \right\} \frac{W_i^{rc\beta-1} e^{-W_i}}{\Gamma(rc\beta) G_{rc\beta} \left(\frac{\beta b_i}{1-\pi_{i1}} \right)} dW_i \left\{ \frac{\prod_{j,k} \phi_{ijk}^{y_{i1jk} + \beta - 1}}{D(\mathbf{y}_i^{(1)} + \beta \mathbf{j})} \right\} \left\{ \frac{\prod_{s=1}^4 \pi_{is}^{y_{is} + \mu_2 \tau_2 - 1}}{D(\mathbf{y}_i^{(2)} + \mu_2 \tau_2)} \right\} d\boldsymbol{\phi}_i^* d\boldsymbol{\pi}_i, \quad (13)$$

with $b_i = \min\{\{1/\phi_{ijk}^*\}, j=1, \dots, r; k=1, \dots, c\}$ and

$$\begin{aligned} \pi_a(\Omega, \mathbf{y}_{(1)} | \mathbf{y}_1, \mathbf{u}, \mathbf{v}, \mathbf{w}) &= \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2} \\ &\quad \prod_{i=1}^A \frac{\Gamma(rc\beta)}{[\Gamma(\beta)]^{rc}} D(\mathbf{y}_i^{(1)} + \beta \mathbf{j}) \\ &\quad \times \prod_{i=1}^A \left(\frac{n_i}{y_{i1}^*, y_{i2}^*, y_{i3}^*, y_{i4}^*} \right) \frac{D(\mathbf{y}_i^{(1)} + \mu_1 \tau_1)}{D(\mu_1 \tau_1)} \frac{D(\mathbf{y}_i^{(2)} + \mu_2 \tau_2)}{D(\mu_2 \tau_2)}. \quad (14) \end{aligned}$$

To evaluate I_i for each $i=1, \dots, A$, we proceed as follows given $(\Omega, \mathbf{y}_{(1)})$:

1. Draw independent samples of vectors $\boldsymbol{\pi}_i$ and $\boldsymbol{\phi}_i^*$ from the Dirichlet($\mathbf{y}_i^{(2)} + \mu_2 \tau_2$) and Dirichlet($\mathbf{y}_i^{(1)} + \beta \mathbf{j}$), respectively. For each $\boldsymbol{\pi}_i$ and $\boldsymbol{\phi}_i^*$, draw a sample of values for W_i from the truncated gamma distribution on the interval $(0, \{\beta b_i / (1 - \pi_{i1})\})$ with parameter $rc\beta$.
2. For each $\boldsymbol{\pi}_i$, $\boldsymbol{\phi}_i^*$ and W_i selected in step (1), compute $R_1 R_2$, where

$$R_1 = G_{rc\beta} \left(\frac{\beta b_i}{1 - \pi_{i1}} \right) / \left[G_{\beta} \left(\frac{\beta}{1 - \pi_{i1}} \right) \right]^{rc} \quad (15)$$

and

$$R_2 = \prod_{j,k} \left\{ \left(\frac{W_i}{\beta} \right) \sum_{s=2}^4 y_{isjk} \left[\frac{1}{\phi_{ijk}^*} \left(1 + \frac{1-\pi_{i1}}{\pi_{i1}} \left\{ 1 - \frac{W_i \phi_{ijk}^*}{\beta} \right\} \right) \right]^{y_{i1jk}} \right\}. \quad (16)$$

3. Repeat steps (1) and (2) 1,000 times. Then compute the average of $R_1 R_2$ over these 1,000 values.

The rest of our computation has two parts. First, we use the griddy Metropolis-Hastings sampler to draw from $\pi_a(\Omega, \mathbf{y}_{(1)} | \mathbf{y}_1, \mathbf{u}, \mathbf{v}, \mathbf{w})$. We sample μ_1, μ_2, τ_1 and τ_2 from their conditional posterior densities using grids; this entails transforming τ_1 and τ_2 to the unit interval $(0, 1)$. For each distribution, 100 grids are used; see Nandram, Cox and Choi (2005) for a similar procedure. Here $\mathbf{y}_{(1)}$ is drawn by sampling from its conditional probability mass function component wise. Draws are made from the conditional posterior density of β using a Metropolis step in a manner similar to Nandram and Choi (2002a, b). We have performed this algorithm 11,000 times and we allowed a “burn-in” of 1,000 iterates. We found that the autocorrelations among the iterates was small, thereby indicating strong mixing of the sampler. We have also used the batch-means method to further assess the computation. We used batches of 25 to compute numerical standard errors.

Second, we use the SIR algorithm to subsample the sample of 10,000 iterates we obtained from $\pi_a(\Omega, \mathbf{y}_{(1)} | \mathbf{y}_1, \mathbf{u}, \mathbf{v}, \mathbf{w})$. For each of the 10,000 iterates we calculate the weights

$$w_m = \frac{h(\Omega^{(m)}, \mathbf{y}_{(1)}^{(m)} | \mathbf{y}_1, \mathbf{u}, \mathbf{v}, \mathbf{w})}{\pi_a(\Omega^{(m)}, \mathbf{y}_{(1)}^{(m)} | \mathbf{y}_1, \mathbf{u}, \mathbf{v}, \mathbf{w})}, \quad m=1, \dots, M=10,000, \quad (17)$$

and we resample $\{\Omega^{(m)}, \mathbf{y}_{(1)}^{(m)}\}$ with probabilities proportional to the weights w_m for $m=1, \dots, M$ without replacement. We use a 10% sampling, and we subsample the 10,000 iterates to get 1,000 iterates; sampling without replacement is a good idea because it avoids repeated values which already exist because the Metropolis-Hastings sampler is not really an accept-reject sampler and it gives repeated values. As usual with sampling without replacement the weights are calculated every time a value is selected.

Finally, we can now make exact (within limitations of Markov chain Monte Carlo methods) inference about \mathbf{p}_i a posteriori. Letting $y_{i,jk} = \sum_{s=1}^4 y_{isjk}$ and \mathbf{y}_i^* denote the vector of $y_{i,jk}$. Then,

$$\mathbf{p}_i | \mathbf{y}_i^*, \mu_1, \tau_1 \sim \text{Dirichlet}(\mathbf{y}_i^* + \mu_1 \tau_1), i=1, \dots, A.$$

Thus, for each value of y_i^* , μ_i , and τ_i we obtain from the SIR algorithm, we draw a value of $p_i, i = 1, \dots, A$. Thus, we obtain a Rao-Blackwellized density for each of the p_i , and inference proceeds in the usual way.

4. An illustrative example

Our illustrative example is in health statistics. In Section 4.1 we briefly discuss the data we used from the third National Health and Nutrition Examination Survey (NHANES III). Specifically, we study the relationship between bone mineral density and family income; see Nandram, Cox and Choi (2005) for a discussion of this problem. In Section 4.2, after briefly discussing our computation, we present posterior inference on the cell probabilities. In section 4.3 using the Bayes factor we discuss the relation between BMD and FI.

4.1 NHANES III data

The sample design is a stratified multistage probability design which is representative of the total civilian non-institutionalized population, 2 months of age or older, in the United States. Further details of the NHANES III sample design are available (National Center for Health Statistics 1992, 1994). The NHANES III data collection consists of two parts: the first part is the sample selection and the interview of the members of a sampled household for their personal information, and the second part is the examination of those interviewed at the mobile examination center (MEC). The health examination has information on physical examination, tests and measurements performed by technicians, and specimen collection. The sample was selected from households in 81 primary units across the continental United States during the period from October 1988 through September 1994. The final data for this study is a part of the 35 largest primary sampling units with population at least 500,000, and we consider 13 subnational areas.

Nonresponse occurs in the interview and examination parts of the survey. The interview nonresponse arises from sampled persons who did not respond for the interview. Some of those who were already interviewed and included in the subsample for a health examination missed the examination at home or at the MEC, thereby missing all or part of the examinations.

Doctors believe that obese and overweight individuals do not generally turn up at the MEC. Cohen and Duffy (2002) point out that “Health surveys are a good example, where it seems plausible that propensity to respond may be related to health.” NHANES III is a good example.

Sampled persons in NHANES III can be categorized by many types of attributes, and researchers analyze such categorical tables for goodness of fit or independence. Here we study bone mineral density (BMD) and family income (FI). We note here that while FI is a discrete variable, we have classified BMD into three levels (normal, osteopenia and osteoporosis), and FI into three levels (low, medium and high). However, only partial classification of the individuals is available because some individuals are classified by only one attribute while others are not classified. About 62% of the households have both FI and BMD observed, 8% with only BMD observed, 29% with only income observed, 1% with neither income nor BMD among those participated in the examination stage. Our problem is to estimate the cell probabilities and to test for association between BMD and FI for each of 13 subnational areas using our expansion model that pools the data adaptively.

In Table 1 we present the 3×3 tables of BMD and FI for the aforementioned 13 areas. Note that areas 6 and 48 have enough data so that they can stand by themselves. However, the other areas are very small; the counts in the table with row totals are generally small except for area 17 and the counts in the table with just total are small. Even for the table with complete data the cell counts are generally small forcing us to use small area estimation techniques to borrow strength.

Table 1
Counts of the 3×3 tables of BMD and FI corresponding to 13 subnational areas in NHANES III

State	Complete Table									Column Total			Row Total			Total
4	21	14	9	8	7	3	2	2	0	11	5	6	4	0	1	1
6	257	127	106	92	51	32	32	5	7	178	54	82	65	28	4	20
12	33	18	21	22	4	4	15	5	0	18	11	16	5	6	2	1
17	25	15	13	8	5	3	0	0	1	18	10	16	17	2	2	4
25	9	7	12	6	5	9	2	1	0	9	6	12	1	4	5	1
26	18	11	18	6	5	9	2	1	1	10	5	11	4	3	0	1
29	9	4	10	3	2	4	3	1	2	9	2	9	0	2	4	1
36	42	17	27	32	13	18	9	6	1	43	21	42	9	7	6	1
39	8	6	14	2	5	4	3	0	1	9	7	5	2	3	0	0
42	14	8	11	12	8	4	8	1	2	35	15	24	3	1	0	0
44	12	9	6	8	5	0	5	1	0	19	4	12	7	1	0	1
48	159	44	22	51	11	13	9	6	2	88	12	23	16	8	2	14
53	14	10	15	10	14	3	1	1		9	4	8	2	4	1	0

Note: In the complete 3×3 table the first (second, third) set of three numbers is the first (second, third) row; the column (row) total refers to the 3×3 table with only column (row) totals; the total refers to the 3×3 table with only total.

4.2 Posterior inference of the cell probabilities

We discuss the performance of our computations for the expansion model, and then we discuss posterior inference about the cell probabilities. We use the posterior mean (PM), posterior standard deviation (PSD) and 95% credible interval for each of parameters of interest. We also present the numerical standard errors (NSE) to assess the repeatability of our computations.

In Table 2 we present summaries of the posterior distributions of $\mu_1, \mu_2, \tau_1, \tau_2$ and β , both before and after the application of the SIR algorithm. These summaries are very similar indicating that the SIR approximation $\pi_{\alpha}(\Omega, y_{(1)} | y_1, u, v, w)$ is not unreasonable. For example, a 95% credible interval for β before and after the application of the SIR algorithm are respectively (1.081, 1.940) and (1.086, 1.947), very good agreement. The estimates of τ_1 and τ_2 should show the largest discrepancies, but these are also reasonably close [e.g., for τ_1 95% credible intervals with the approximation is (28.282, 64.204) and with the SIR algorithm it is (27.962, 64.425)]. In both cases the NSEs are small indicating that the computations are repeatable.

In Table 3 we have compared our expansion model (Model 3) with two other models. Model 1, an ignorable nonresponse model, and Model 2, a nonignorable nonresponse model (no centering), are described in Appendix B. For illustration we have selected three areas, a large area, a medium-sized area and a small area. There are differences among the three models. In general, the larger estimates tend to be smaller for Model 2, and even smaller than Model

1, than for Model 3 (i.e., the estimates from Model 3 are naturally closest to Model 1, and not Model 2). Model 2 produces the largest variability; as expected, Model 3 gives slightly larger variability than Model 1. Because of space restrictions we have not presented the NSEs, but we note that they are all smaller than 0.005.

4.3 Bayes factor for evidence of association

We have also considered the association between BMD and FI. It appears doubtful whether such an association might exist, but it is interesting to look at this issue; see Nandram, Cox and Choi (2005) for a discussion on this problem. We use the Bayes factor (Kass and Raftery 1995) to measure the strength of the evidence of an association relative to no association in the $r \times c$ categorical table. We have done so for each of the thirteen areas and all areas combined.

We have used two procedures, one without extensive modeling and the other using our nonignorable nonresponse (expansion) model. The simple method is to fill in the cell counts using an ordinary raking procedure, and we assume there is no error in doing so. This is a common sense procedure that survey practitioners have used routinely. In the second procedure using our nonignorable nonresponse model, we have obtained 1,000 combined tables for each area as described in Section 3 on computations. For each area we have obtained the cell counts for all four tables, and we summed them to get a single table of all counts.

Table 2

NHANES data on 13 areas: Comparison of the approximate posterior density and the correct posterior density using the posterior means (PM), posterior standard deviations (PSD), numerical standard errors (NSE) and 95% credible intervals of the hyperparameters

	Approximation				Adjusted			
	PM	PSD	NSE	95% Int	PM	PSD	NSE	95% Int
μ_{21}	0.528	0.031	0.001	(0.463, 0.582)	0.525	0.031	0.008	(0.456, 0.578)
μ_{22}	0.131	0.021	0.001	(0.096, 0.181)	0.133	0.021	0.002	(0.094, 0.179)
μ_{23}	0.328	0.028	0.001	(0.274, 0.383)	0.328	0.028	0.005	(0.269, 0.383)
μ_{24}	0.013	0.006	0.000	(0.004, 0.027)	0.014	0.006	0.000	(0.004, 0.029)
τ_2	21.638	9.559	0.255	(8.347, 46.587)	20.078	8.632	0.303	(8.538, 38.625)
μ_{111}	0.280	0.023	0.001	(0.234, 0.324)	0.277	0.023	0.004	(0.228, 0.319)
μ_{112}	0.133	0.016	0.000	(0.102, 0.165)	0.134	0.017	0.002	(0.101, 0.165)
μ_{113}	0.200	0.019	0.000	(0.163, 0.238)	0.199	0.019	0.003	(0.162, 0.236)
μ_{121}	0.105	0.015	0.000	(0.078, 0.135)	0.107	0.015	0.002	(0.079, 0.135)
μ_{122}	0.065	0.011	0.000	(0.044, 0.088)	0.065	0.011	0.001	(0.044, 0.087)
μ_{123}	0.072	0.012	0.000	(0.050, 0.096)	0.073	0.012	0.001	(0.049, 0.097)
μ_{131}	0.061	0.011	0.000	(0.041, 0.083)	0.061	0.011	0.001	(0.040, 0.083)
μ_{132}	0.037	0.008	0.000	(0.023, 0.054)	0.036	0.008	0.001	(0.022, 0.054)
μ_{133}	0.048	0.009	0.000	(0.031, 0.068)	0.048	0.009	0.001	(0.031, 0.068)
τ_1	45.960	10.094	0.153	(28.282, 64.204)	45.177	10.562	0.679	(27.962, 64.423)
β	1.472	0.218	0.004	(1.081, 1.940)	1.449	0.208	0.022	(1.086, 1.947)

Note: The hyperparameters are $\mu_1, \mu_2, \tau_1, \tau_2$ and β .

Table 3
Posterior means of the cell probabilities and 95% credible intervals (CI) for three areas (large, medium and small) by the three models

Cell	Model 1			Model 2			Model 3		
	PM	PSD	95% CI	PM	PSD	95% CI	PM	PSD	95% CI
a. Large									
(1,1)	0.239	0.044	(0.157, 0.326)	0.196	0.046	(0.117, 0.295)	0.259	0.038	(0.189, 0.335)
(1,2)	0.140	0.035	(0.078, 0.213)	0.127	0.035	(0.068, 0.200)	0.132	0.029	(0.082, 0.197)
(1,3)	0.240	0.044	(0.159, 0.332)	0.198	0.047	(0.118, 0.301)	0.248	0.037	(0.175, 0.322)
(2,1)	0.092	0.032	(0.039, 0.162)	0.098	0.040	(0.037, 0.188)	0.077	0.022	(0.039, 0.126)
(2,2)	0.074	0.028	(0.029, 0.136)	0.077	0.030	(0.030, 0.144)	0.056	0.020	(0.024, 0.099)
(2,3)	0.133	0.036	(0.070, 0.210)	0.121	0.042	(0.056, 0.219)	0.110	0.028	(0.058, 0.168)
(3,1)	0.036	0.020	(0.008, 0.083)	0.069	0.039	(0.013, 0.153)	0.047	0.018	(0.018, 0.086)
(3,2)	0.023	0.015	(0.003, 0.061)	0.043	0.025	(0.007, 0.100)	0.032	0.014	(0.009, 0.063)
(3,3)	0.025	0.017	(0.003, 0.066)	0.071	0.040	(0.010, 0.154)	0.042	0.016	(0.016, 0.079)
b. Medium									
(1,1)	0.233	0.034	(0.169, 0.302)	0.213	0.043	(0.141, 0.305)	0.254	0.032	(0.194, 0.318)
(1,2)	0.143	0.028	(0.093, 0.200)	0.127	0.032	(0.072, 0.196)	0.146	0.024	(0.102, 0.197)
(1,3)	0.190	0.031	(0.132, 0.254)	0.140	0.034	(0.084, 0.218)	0.208	0.027	(0.156, 0.259)
(2,1)	0.174	0.031	(0.118, 0.237)	0.160	0.042	(0.092, 0.249)	0.154	0.027	(0.106, 0.211)
(2,2)	0.043	0.018	(0.015, 0.083)	0.060	0.028	(0.017, 0.124)	0.032	0.012	(0.012, 0.059)
(2,3)	0.049	0.020	(0.017, 0.095)	0.065	0.031	(0.018, 0.136)	0.042	0.014	(0.020, 0.072)
(3,1)	0.112	0.025	(0.068, 0.167)	0.120	0.041	(0.059, 0.209)	0.092	0.020	(0.056, 0.134)
(3,2)	0.047	0.018	(0.018, 0.088)	0.059	0.026	(0.019, 0.118)	0.040	0.014	(0.018, 0.071)
(3,3)	0.010	0.009	(0.000, 0.033)	0.056	0.032	(0.006, 0.122)	0.032	0.012	(0.013, 0.059)
c. Small									
(1,1)	0.196	0.052	(0.103, 0.305)	0.164	0.055	(0.077, 0.288)	0.253	0.043	(0.175, 0.334)
(1,2)	0.081	0.034	(0.028, 0.158)	0.081	0.032	(0.030, 0.155)	0.091	0.028	(0.043, 0.152)
(1,3)	0.213	0.052	(0.118, 0.323)	0.175	0.055	(0.087, 0.300)	0.220	0.043	(0.137, 0.306)
(2,1)	0.093	0.041	(0.028, 0.186)	0.111	0.055	(0.029, 0.234)	0.073	0.028	(0.030, 0.139)
(2,2)	0.056	0.029	(0.012, 0.126)	0.066	0.031	(0.018, 0.136)	0.045	0.020	(0.014, 0.094)
(2,3)	0.115	0.045	(0.042, 0.215)	0.118	0.053	(0.038, 0.240)	0.092	0.030	(0.041, 0.158)
(3,1)	0.115	0.048	(0.036, 0.222)	0.113	0.056	(0.031, 0.239)	0.081	0.030	(0.033, 0.148)
(3,2)	0.044	0.028	(0.006, 0.113)	0.065	0.035	(0.013, 0.144)	0.043	0.020	(0.012, 0.086)
(3,3)	0.087	0.042	(0.022, 0.184)	0.107	0.055	(0.023, 0.227)	0.103	0.034	(0.047, 0.181)

Note: See Appendix B for a description of Models 1 and 2.

The raking procedure to obtain the cell counts is described as follows. Let n_{jk} denote the cell counts for the four tables combined. Let $n_{jk}^{(1)}$ denote the cell counts for the table with complete data, $n_{j,c+1}^{(2)}$ denote the table with row totals, $n_{r+1,k}^{(3)}$ denote the table with column totals and $n_{r+1,c+1}^{(4)}$ denote the table with total. The cell counts for the four tables are estimated as

$$n_{jk} = n_{jk}^{(1)} + \left(\frac{n_{jk}^{(1)}}{n_{j\cdot}^{(1)}}\right)n_{j,c+1}^{(2)} + \left(\frac{n_{jk}^{(1)}}{n_{\cdot k}^{(1)}}\right)n_{r+1,k}^{(3)} + \left(\frac{n_{jk}^{(1)}}{n_{\cdot}^{(1)}}\right)n_{r+1,c+1}^{(4)},$$

$$j = 1, \dots, r, k = 1, \dots, c.$$

In either case we denote the sum of the cell counts for each area by $n_{jk\cdot}$. For the raking procedure we have a single table for each area, and for the nonignorable nonresponse model we have a sample of 1,000 tables for each area. We also have a single combined table for all areas under the raking procedure and 1,000 tables for all areas combined. We obtain the Bayes factor for each table under a multinomial-Dirichlet model. It is worth noting that our method uses the expansion model so that the cell counts borrow strength from other areas unlike the raking procedure.

Then, for each table we take

$$\mathbf{n} | \boldsymbol{\pi} \sim \text{Multinomial}(n, \boldsymbol{\pi}) \text{ and } \boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{1}).$$

That is, we take a uniform prior for $\boldsymbol{\pi}$ with $\pi_{jk} > 0$ and $\sum_{j=1}^r \sum_{k=1}^c \pi_{jk} = 1$. Under the hypothesis of no association we have $\pi_{jk} = \pi_j \pi_k$, where $\pi_j > 0$, $\sum_{j=1}^r \pi_j = 1$ and $\pi_k > 0$, $\sum_{k=1}^c \pi_k = 1$. Thus, the hypothesis of association is that the π_{jk} are unrestricted (except that they are nonnegative and sum to unity) whereas for the hypothesis of no associate $\pi_{jk} = \pi_j \pi_k$.

The Bayes factor is the ratio of the marginal likelihood under association versus the marginal likelihood under no association. This measures the strength of evidence of association versus no association; see Kass and Raftery (1995). Let $p_a(\mathbf{n})$ denote the marginal likelihood under association and $p_0(\mathbf{n})$ denote the marginal likelihood under no association. Then, letting $n_{j\cdot} = \sum_{k=1}^c n_{jk}$ and $n_{\cdot k} = \sum_{j=1}^r n_{jk}$, it is easy to show that

$$p_a(\mathbf{n}) = p_0(\mathbf{n}) \left\{ \prod_{u=0}^{n-1} \frac{u + rc}{(u + r)(u + c)} \right\} \frac{\prod_{j=1}^r n_{j\cdot}! \prod_{k=1}^c n_{\cdot k}!}{\prod_{j=1}^r \prod_{k=1}^c n_{jk}!},$$

where $p_0(\mathbf{n}) = n! \prod_{u=0}^{n-1} (u+rc)^{-1}$. Observe that $p_0(\mathbf{n})$ is not a function of $\{n_{jk}\}$. Thus, as a measure of association it is the deviation of $\prod_{j=1}^r n_{j\cdot}! \prod_{k=1}^c n_{\cdot k}!$ from $\prod_{j=1}^r \prod_{k=1}^c n_{jk}!$ that matters. However, we note that for the classical Pearson statistic for independence it is the deviations of n_{jk} from $n_{j\cdot}n_{\cdot k}$ that matter. But note that this test cannot be applied because many of the expected cell counts are smaller than 5 under the hypothesis of no association and multinomial sampling.

We present our results in Table 4 and in Figure 1 corresponding to the data in Table 1 for the cross-classification of BMD and FI. We have presented the

logarithms of the marginal likelihoods (base e) and the Bayes factors; these are to be interpreted using the rule of thumb of Kass and Raftery (1995).

In Figure 1 we can see that the box plots are all above zero except the one for the third area which provides no evidence for association; perhaps there is no evidence for association in area 42 (10 in figure) as well. A summary of these results are presented in Table 4. The Bayes factors show association in all areas, except area 12, and they are much larger under the nonignorable nonresponse model. Area 6 and all areas combined are elevated (336.3 vs. 5.8 and 3,798.2 vs. 0.183).

Table 4
NHANES data on 13 areas: Comparison of the negative marginal likelihoods and Bayes factors of association of BMD and FI from the raking procedure and the expansion model by area

area	$-\ln\{p_0(\mathbf{n})\}$	Raking		Expansion	
		$-\ln\{p_a(\mathbf{n})\}$	BF	$-\ln\{p_a(\mathbf{n})\}$	BF
4	26.19	23.07	22.855	23.5 _{0.014}	14.78 _{0.169}
6	45.73	43.98	5.766	40.5 _{0.038}	336.27 _{11.465}
12	31.14	38.01	0.001	33.4 _{0.054}	0.37 _{0.027}
17	29.13	27.03	8.134	27.0 _{0.026}	10.27 _{0.191}
25	25.44	26.02	0.558	23.8 _{0.029}	9.55 _{0.202}
26	26.89	23.18	40.562	23.9 _{0.018}	24.71 _{0.370}
29	23.21	20.87	10.301	21.3 _{0.018}	8.40 _{0.115}
36	34.99	36.09	0.330	33.1 _{0.064}	21.13 _{0.928}
39	23.77	24.89	0.325	23.6 _{0.044}	2.24 _{0.68}
42	29.51	30.21	0.497	30.3 _{0.099}	4.33 _{0.255}
44	25.61	30.48	0.008	24.4 _{0.027}	5.19 _{0.137}
48	38.83	35.34	32.650	39.1 _{0.060}	2.15 _{0.081}
53	27.11	24.82	9.865	24.2 _{0.017}	19.40 _{0.282}
All	53.43	55.13	0.183	46.1 _{0.049}	3,798.24 _{51.82}

Note: Area 'all' refers to all areas combined; the notation a_b means that the average is a and the standard error is b over the 1,000 iterates; $\ln\{p_0(\mathbf{n})\}$ is the same for both procedures.

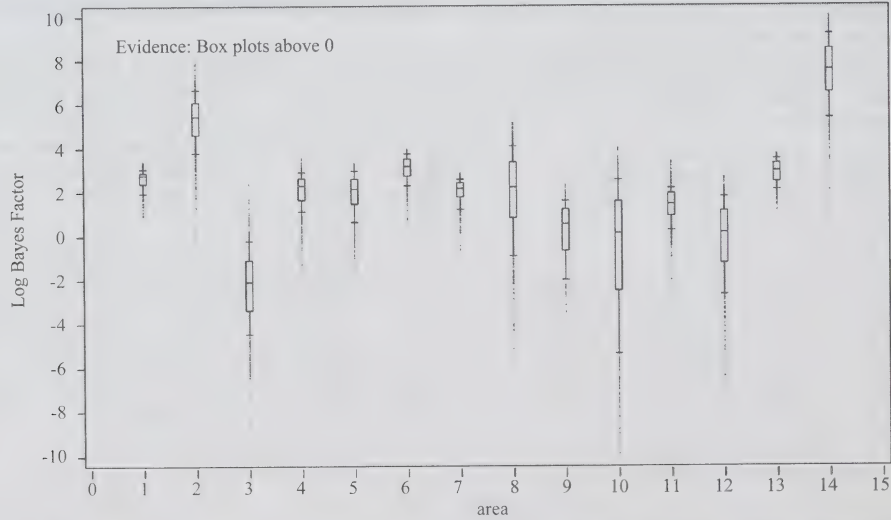


Figure 1 Box plots of log Bayes factors versus areas to measure evidence for association between BMD and FI

5. Concluding remarks

The purpose of this paper has been to develop a methodology to analyze data from incomplete two-way categorical tables, each table corresponding to an area. We have done so by extending the Bayesian methodology of Nandram and Choi (2002a, b) for binary data to $r \times c$ categorical tables for small areas. We have constructed a new Bayesian nonignorable nonresponse model (*i.e.*, expansion model) which is centered on the ignorable nonresponse model. We have used Markov chain Monte Carlo methods (specifically the gridy Metropolis-Hastings sampler) to fit the model. We have compared our model with an ignorable nonresponse model and a nonignorable nonresponse model. Finally, we have done an illustrative example on estimating the cell probabilities for the 3×3 table of BMD and income over thirteen subnational areas.

We have shown that there are differences among the three models. Using the data on BMD and FI, we have shown that our expansion model is a compromise between the ignorable nonresponse model and the nonignorable nonresponse model. Using the Bayes factor we have shown that there are differences between the tests of association for BMI and FI when the cell counts are estimated from our model and when using a raking procedure. In fact, owing to the borrowing of strength, we have seen that the evidence for association under our model is much stronger than the from the raking procedure.

There are three additional avenues that we can explore. First, we can construct a model to incorporate systematic departure from ignorability. This task will need more costly field work to get the much-needed information. Second, it is also interesting to relax the assumption that the margins of the categorical table are fixed; see, for example, Nandram (2009) who looked at a single large area. Third, there can be further improvement in calibration (*i.e.*, incorporating information about margins).

Acknowledgements

We are grateful to the two referees and the associate editor for their assistance in the presentation of the material.

Appendix A

Joint posterior density of the expansion model

First, integrating the joint posterior density over \boldsymbol{p} we get that

$$\begin{aligned}
 h(\Omega, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{y}_{(1)} | \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) &\propto \prod_{i=1}^4 \frac{D(\boldsymbol{y}_i^{(1)} + \boldsymbol{\mu}_1 \boldsymbol{\tau}_1)}{D(\boldsymbol{\mu}_1 \boldsymbol{\tau}_1)} \\
 &\times \left[\binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} (1 - \pi_{i1})^{-y_i} \right. \\
 &\times \left\{ \prod_{s=1}^4 \pi_{is}^{y_{is}} \prod_{j,k} \phi_{ijk}^{y_{i,jk}} [a_{ijk}(\pi_{i1}, \phi_{ijk})]^{y_{i,jk}} \right\} \\
 &\times \left\{ \frac{\prod_{s=1}^4 \pi_{is}^{u_{2s} \tau_2 - 1}}{D(\boldsymbol{\mu}_2 \boldsymbol{\tau}_2)} \prod_{j=1}^r \prod_{k=1}^c \frac{\beta^\beta}{(1 - \pi_{i1})^\beta} \right. \\
 &\left. \left. \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta \phi_{ijk}}{1 - \pi_{i1}}}}{\Gamma(\beta) G_\beta [\beta(1 - \pi_{i1})^{-1}]} \right\} \right] \\
 &\times \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2}, \tag{A.1}
 \end{aligned}$$

where the $rc \times 1$ vector

$$\boldsymbol{y}_i^{(1)} \stackrel{\text{def}}{=} (y_{i,11}, y_{i,12}, \dots, y_{i,1c}, y_{i,21}, \dots, y_{i,2c}, y_{i,r1}, \dots, y_{i,rc})'.$$

Now let \boldsymbol{j} denote an $rc \times 1$ vector of ones and let

$$\boldsymbol{y}_i^{(2)} \stackrel{\text{def}}{=} (y_{i1..}, y_{i2..}, y_{i3..}, y_{i4..})'.$$

Then, collapsing over $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$, we have that

$$\begin{aligned}
 h(\Omega, \boldsymbol{y}_{(1)} | \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) &\propto \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2} \\
 &\times \prod_{i=1}^A \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \frac{D(\boldsymbol{y}_i^{(1)} + \boldsymbol{\mu}_1 \boldsymbol{\tau}_1)}{D(\boldsymbol{\mu}_1 \boldsymbol{\tau}_1)} \frac{D(\boldsymbol{y}_i^{(2)} + \boldsymbol{\mu}_2 \boldsymbol{\tau}_2)}{D(\boldsymbol{\mu}_2 \boldsymbol{\tau}_2)} \\
 &\times \prod_{i=1}^A \left[\frac{\beta^\beta}{\Gamma(\beta)} \right]^{\tau_c} D(\boldsymbol{y}_i^{(1)} + \beta \boldsymbol{j}) I_i, \tag{A.2}
 \end{aligned}$$

where

$$\begin{aligned}
 I_i &= \int \prod_{j,k} \left[\frac{\left[\left(\frac{1 - \pi_{i1}}{\phi_{ijk}} \right) \left(1 + \frac{1}{\pi_{i1}} \{ 1 - \pi_{i1} - \phi_{ijk} \} \right) \right]^{y_{i,jk}}}{(1 - \pi_{i1})^{y_{i,jk} + \beta} G_\beta \left(\frac{\beta}{1 - \pi_{i1}} \right)} \right] \\
 &\times \left\{ \prod_{j,k} e^{-\frac{\beta \phi_{ijk}}{1 - \pi_{i1}}} \right\} \left\{ \frac{\prod_{j,k} \phi_{ijk}^{y_{i,jk} + \beta - 1}}{D(\boldsymbol{y}_i^{(1)} + \beta \boldsymbol{j})} \right\} \\
 &\times \left\{ \frac{\prod_{s=1}^4 \pi_{is}^{y_{is} + \mu_{2s} \tau_2 - 1}}{D(\boldsymbol{y}_i^{(2)} + \boldsymbol{\mu}_2 \boldsymbol{\tau}_2)} \right\} d\boldsymbol{\phi} d\boldsymbol{\pi}_i. \tag{A.3}
 \end{aligned}$$

Note that $0 \leq \pi_{is} \leq 1, \sum_{s=1}^4 \pi_{is} = 1$ and $0 \leq \phi_{ijk} \leq 1$. We simplify the computation for I_i in (A.3) in two steps.

First, in (A.3) we make the transformation

$$\phi_{ijk} = T_i \phi_{ijk}^* \quad \sum_{j=1}^r \sum_{k=1}^c \phi_{ijk} = T_i.$$

The new variables ϕ_{ijk}^* satisfy the relationships $0 \leq \phi_{ijk}^* \leq 1, \sum_{j=1}^r \sum_{k=1}^c \phi_{ijk}^* = 1$ and the T_i are restricted so that $0 \leq T_i \leq 1 / \phi_{ijk}^*$, for $j = 1, \dots, r, k = 1, \dots, c$ and $0 \leq T_i \leq rc$. With this transformation we have

$$I_i = \iiint_0^1 \prod_{j,k} \left[\frac{T_i}{1 - \pi_{i1}} \right]^{\sum_{s=2}^4 y_{isjk}} \left[\frac{1}{\phi_{ijk}^*} \left(1 + \frac{1 - \pi_{i1}}{\pi_{i1}} \left\{ 1 - \frac{T_i}{1 - \pi_{i1}} \phi_{ijk}^* \right\} \right) \right]^{y_{i1jk}} \times \left\{ \frac{\left(\frac{T_i}{1 - \pi_{i1}} \right)^{rc\beta-1} e^{-\frac{\beta T_i}{1 - \pi_{i1}}}}{(1 - \pi_{i1}) \left[G_\beta \left(\frac{\beta}{1 - \pi_{i1}} \right) \right]^{rc}} \right\} \left\{ \frac{\prod_{j,k} \phi_{ijk}^{y_{i,jk} + \beta - 1}}{D(y_i^{(1)} + \beta j)} \right\} \times \left\{ \frac{\prod_{s=1}^4 \pi_{is}^{y_{is} + \mu_{2s} \tau_2 - 1}}{D(y_i^{(2)} + \mu_2 \tau_2)} \right\} dT_i d\phi_{ijk}^* d\pi_i,$$

where $b_i = \min \{1 / \phi_{ijk}^* | j = 1, \dots, r; k = 1, \dots, c\}$.

Second, letting $W_i = \{\beta T_i / 1 - \pi_{i1}\}$ and absorbing the factor $\beta^{rc\beta} / \Gamma(rc\beta)$ in I_i , with some additional algebra, we have

$$I_i = \iiint_0^1 \prod_{j,k} \left[\frac{\beta b_i}{1 - \pi_{i1}} \frac{G_{rc\beta} \left(\frac{\beta b_i}{1 - \pi_{i1}} \right)}{\left[G_\beta \left(\frac{\beta}{1 - \pi_{i1}} \right) \right]^{rc}} \right] \prod_{j,k} \left\{ \left(\frac{W_i}{\beta} \right)^{\sum_{s=2}^4 y_{isjk}} \left[\frac{1}{\phi_{ijk}^*} \left(1 + \frac{1 - \pi_{i1}}{\pi_{i1}} \left\{ 1 - \frac{W_i \phi_{ijk}^*}{\beta} \right\} \right) \right]^{y_{i1jk}} \right\} \times \frac{W_i^{rc\beta-1} e^{-W_i}}{\Gamma(rc\beta) G_{rc\beta} \left(\frac{\beta b_i}{1 - \pi_{i1}} \right)} \left\{ \frac{\prod_{j,k} \phi_{ijk}^{y_{i,jk} + \beta - 1}}{D(y_i^{(1)} + \beta j)} \right\} \left\{ \frac{\prod_{s=1}^4 \pi_{is}^{y_{is} + \mu_{2s} \tau_2 - 1}}{D(y_i^{(2)} + \mu_2 \tau_2)} \right\} d\phi_{ijk}^* d\pi_i. \quad (A.4)$$

Appendix B

Ignorable and nonignorable nonresponse models

Set $\psi_{ijk} \equiv 1$ in the expansion model to form the ignorable nonresponse model. For $i = 1, \dots, A$, we then take

$$\pi_i | \mu_2, \tau_2 \sim \text{Dirichlet}(\mu_2 \tau_2)$$

and independently

$$p_i | \mu_1, \tau_1 \sim \text{Dirichlet}(\mu_1 \tau_1).$$

Also, $p(\tau_2) = \{1 / (1 + \tau_1)^2\}$, $\tau_1 \geq 0, \mu_1 \sim \text{Dirichlet}(1)$, $p(\tau_2) = \{1 / (1 + \tau_2)^2\}$, $\tau_1 \geq 0$ and $\mu_2 \sim \text{Dirichlet}(1)$. Here we have independence at all levels and the vectors $\mathbf{1}$ are of the appropriate dimension with every coordinate equal to one. Note, that all the parameters of the ignorable model are identifiable and estimable.

Set $\pi_{isjk} = \pi_{is} \psi_{ijk}$ in the expansion model to form the nonignorable nonresponse model. In this case, for $i = 1, \dots, A$,

$$\pi_{ijk} | \mu_2, \tau_2 \sim \text{Dirichlet}(\mu_2 \tau_2)$$

and independently

$$p_i | \mu_1, \tau_1 \sim \text{Dirichlet}(\mu_1 \tau_1).$$

In this model, the parameters π_{ijk} are not identifiable and we take $\tau_2 \sim \text{Gamma}(\alpha_0, \beta_0)$, where α_0 and β_0 are to be specified. The model specification is then completed by assigning τ_1, μ_1 and μ_2 the same distributional properties as in the previous paragraph.

As in Nandram, Cox and Choi (2005), α_0 and β_0 are specified as follows. The ignorable nonresponse model is fit to obtain a sample from the posterior density of τ_2 . Then α_0 and β_0 are obtained using the method of moments. Nandram, Cox and Choi (2005) found that inference about p_i is not very sensitive to the choice of these parameters.

References

- Cohen, G., and Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics*, 18, 13-23.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, 57, 45-97.
- Forster, J.J., and Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society, Series B*, 60, 57-70.
- Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

- Greenland, S. (2009). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Sciences*, 24, 195-210.
- Kass, R.E., and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Second Edition. New York: John Wiley & Sons, Inc.
- Nandram, B. (2009). Bayesian inference of the cell probabilities of a two-way categorical table under non-ignorability. *Communications in Statistics - Theory and Methods*, 38, 3015-3030.
- Nandram, B., and Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.
- Nandram, B., and Choi, J.W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.
- Nandram, B., and Choi, J.W. (2004). A nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse. *Journal of Nonparametric Statistics*, 16, 821-839.
- Nandram, B., and Choi, J.W. (2005). Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the NHANES data. *Survey Methodology*, 31, 73-84.
- Nandram, B., and Choi, J.W. (2008). A Bayesian allocation of undecided voters. *Survey Methodology*, 34, 37-49.
- Nandram, B., and Choi, J.W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, 105, 120-135.
- Nandram, B., Cox, L.H. and Choi, J.W. (2005). Bayesian analysis of nonignorable missing categorical data: An application to bone mineral density and family income. *Survey Methodology*, 31, 213-225.
- Nandram, B., Han, G. and Choi, J.W. (2002). A hierarchical Bayesian nonignorable nonresponse model for multinomial data from small areas. *Survey Methodology*, 28, 145-156.
- Nandram, B., Liu, N., Choi, J.W. and Cox, L. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.
- National Center for Health Statistics (1992). Third national health and nutrition examination survey. *Vital and Health Statistics*, Series 2, 113.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Smith, A.F.M., and Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46, 84-88.

Why one should incorporate the design weights when adjusting for unit nonresponse using response homogeneity groups

Phillip S. Kott¹

Abstract

When there is unit (whole-element) nonresponse in a survey sample drawn using probability-sampling principles, a common practice is to divide the sample into mutually exclusive groups in such a way that it is reasonable to assume that each sampled element in a group were equally likely to be a survey nonrespondent. In this way, unit response can be treated as an additional phase of probability sampling with the inverse of the estimated probability of unit response within a group serving as an adjustment factor when computing the final weights for the group's respondents. If the goal is to estimate the population mean of a survey variable that roughly behaves as if it were a random variable with a constant mean within each group regardless of the original design weights, then incorporating the design weights into the adjustment factors will usually be more efficient than not incorporating them. In fact, if the survey variable behaved exactly like such a random variable, then the estimated population mean computed with the design-weighted adjustment factors would be nearly unbiased in some sense (*i.e.*, under the combination of the original probability-sampling mechanism and a prediction model) even when the sampled elements within a group are not equally likely to respond.

Key Words: Double protection; Prediction model; Probability sampling; Response model; Sampling phase; Stratified Bernoulli sampling.

1. Introduction

In the absence of nonresponse, it is possible to estimate the mean of a finite population from a survey sample without having to assume a statistical model which, no matter how reasonable, may not hold true. This is done by assigning each element of the population a positive probability of sample selection and creating estimators around this random-selection mechanism. Unfortunately, surveys taken in the real world often suffer from nonresponse.

Two different types of models can be used in the face of unit (whole-element) nonresponse. One is a prediction or outcome model in which the survey variable of interest is assumed to behave like a random variable with known characteristics but unknown parameters. The other is a response or selection model where the very act of an element's responding to a survey is treated as an additional phase of random sample selection.

Conventionally, survey statisticians prefer response models for two reasons. In addition to the convenience of response modeling allowing them to treat unit response as an additional phase of random sampling, a survey is usually designed to collect information on a number of variables from the sampled elements. Prediction modeling requires assuming a different model for each survey variable any one of which could fail. Response modeling, by contrast, requires only the assumption of a single model. This is no longer true when there is item (survey-variable-specific) nonresponse. Consequently, prediction modeling is more common when

handling item nonresponse through imputation. That being said, item nonresponse is beyond the scope of this note.

Under an assumed response model, the element probabilities of response are treated as unknown, which means that they have to be estimated from the sample. Typically, the response mechanism is assumed to be independent across elements and not to depend on whether the element is in the sample (each element has an *a priori* probability of response which becomes operational if it is selected for the sample). The simplest and mostly commonly used response model separates the sample, and implicitly the entire population, into mutually exclusive groups, called "response homogeneity groups" by Särndal, Swensson and Wretman (1992) (the term "weighting classes" is more common; see, for example, Lohr (2009, pages 340-341)), and assumes that each element in a group is equally likely to be a unit respondent regardless of its probability of selection into the original sample, π_k . Thus, the response mechanism produces a stratified Bernoulli subsample with the groups serving as the strata.

Conditioned on the respondent sample sizes in the groups, a stratified Bernoulli subsample with unknown selection (response) probabilities is converted into a stratified simple random subsample with known selection probabilities: r_g/n_g for the elements in group g when that group has n_g sampled elements, r_g or which respond.

Although the conditional probabilities of response in group g under the stratified Bernoulli response model is r_g/n_g , we will see it is often better to multiply the design

1. Phillip S. Kott, RTI International, Suite 902, 6100 Executive Blvd., Rockville, MD 20852, U.S.A. E-mail: pkott@rti.org.

weight, $d_k = 1/\pi_k$, for a responding element in the group not by n_g/r_g , but by

$$f_g = \frac{\sum_{k \in S_g} d_k}{\sum_{k \in R_g} d_k}, \quad (1)$$

where S_g is the original sample and R_g the respondent subsample in group g . This *adjustment factor* can be different from n_g/r_g when the d_k in group g vary.

Little and Vartivarian (2003) claim that using the f_g is what is usually done in practice. They argue, however, that incorporating design weights into the adjustment factor in this way can “add to the variance”.

In section 2, we develop the notation for estimating the population mean of a survey variable. Using the n_g/r_g produces a double-expansion estimator, while using the f_g produces a reweighted-expansion estimator. We can express both using a formulation in Kim, Navarro and Fuller (2006). From that expression, it is possible to see that if the survey variable roughly behaves like a random variable with a constant mean within each group regardless of the design weights, then using the f_g will often be more efficient than using the n_g/r_g . In fact, if the survey variable behaved exactly like such a random variable, then the estimated population mean computed with the f_g would be nearly unbiased under the combination of the original sampling design and this prediction model even when the response model fails.

In Section 3, we show that empirical results in Little and Vartivarian (2003) are consistent with these arguments and offer some concluding remarks.

2. The two estimators

Suppose we want to estimate the population mean of a survey variable y_k :

$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{\sum_{g=1}^G \sum_{k \in U_g} y_k}{\sum_{g=1}^G N_g} = \frac{\sum_{g=1}^G N_g \bar{y}_{U_g}}{\sum_{g=1}^G N_g},$$

where the population U is divided into G groups, U_1, \dots, U_G , each U_g contains N_g elements, and $N = N_1 + \dots + N_G$. In the absence of nonresponse, each N_g is estimated in an unbiased fashion under probability-sampling theory by $\hat{N}_g = \sum_{k \in S_g} d_k$, and each \bar{y}_{U_g} is estimated in a nearly (*i.e.*, asymptotically) unbiased fashion

$$\bar{y}_{S_g} = \frac{\sum_{k \in S_g} d_k y_k}{\sum_{k \in S_g} d_k}, \quad (2)$$

under mild conditions when n_g is sufficiently large. We assume both here.

For a formal statement of the conditions under which each \bar{y}_{S_g} is consistent under probability sampling theory and therefore nearly unbiased, see Fuller (2009, page 115). The interested reader is directed to Fuller whenever a result in this note depends on assumptions about the design and population as the sample size grows arbitrarily large. A more rigorous treatment of much of what is to be discussed here under the response model can be found in Kim, Navarro and Fuller (2006).

Let us label the full-sample estimator for \bar{y}_U we have been discussing $\bar{y}_S = \sum_g \hat{N}_g \bar{y}_{S_g}$. There are more direct ways to render \bar{y}_S , but this version will better serve our purposes.

If we adjust for nonresponse using the f_g in equation (1), we have the reweighted-expansion estimator:

$$\begin{aligned} \hat{y}_{rw} &= \frac{\sum_{g=1}^G \left(f_g \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^G \left(f_g \sum_{k \in R_g} d_k \right)} \\ &= \frac{\sum_{g=1}^G \left(\frac{\sum_{k \in S_g} d_k}{\sum_{k \in R_g} d_k} \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^G \left(\frac{\sum_{k \in S_g} d_k}{\sum_{k \in R_g} d_k} \sum_{k \in R_g} d_k \right)} = \frac{\sum_{g=1}^G \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k y_k}{\sum_{k \in R_g} d_k} \right)}{\sum_{g=1}^G \hat{N}_g}. \end{aligned}$$

Technically, \hat{y}_{rw} is the ratio of two reweighted-expansion estimators, but we use the simpler terminology here.

Employing the n_g/r_g results in the double-expansion estimator:

$$\hat{y}_{de} = \frac{\sum_{g=1}^G \left(\frac{n_g}{r_g} \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^G \left(\frac{n_g}{r_g} \sum_{k \in R_g} d_k \right)}.$$

For our purposes, this estimator can also be expressed as

$$\hat{y}_{de} = \frac{\sum_{g=1}^G \left(\frac{\sum_{k \in S_g} d_k p_k}{\sum_{k \in R_g} d_k p_k} \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^G \left(\frac{\sum_{k \in S_g} d_k p_k}{\sum_{k \in R_g} d_k p_k} \sum_{k \in R_g} d_k \right)} = \frac{\sum_{g=1}^G \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k y_k}{\sum_{k \in R_g} d_k p_k} \right)}{\sum_{g=1}^G \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k}{\sum_{k \in R_g} d_k p_k} \right)},$$

where

$$p_k = \frac{1}{d_k} \frac{\sum_{j \in S_g} d_j}{n_g} \quad \text{for } k \in S_g \quad (3)$$

(so that $\sum_{k \in S_g} d_k p_k = \sum_{k \in S_g} d_k = \hat{N}_g$).

Both \hat{y}_{rw} and \hat{y}_{de} can now be written in the form:

$$\hat{y}_{S,q} = \frac{\sum_{g=1}^G \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k y_k}{\sum_{k \in R_g} d_k q_k} \right)}{\sum_{g=1}^G \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k}{\sum_{k \in R_g} d_k q_k} \right)}. \quad (4)$$

For the reweighted-expansion estimator, all $q_k = 1$, while for the double-expansion estimator, $q_k = p_k$ as defined by equation (3).

We will soon have use of the following for our two estimators:

$$\hat{y}_{S,q} - \bar{y}_S = \frac{\sum_{g=1}^G \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k e_k}{\sum_{k \in R_g} d_k q_k} \right)}{\sum_{g=1}^G \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k}{\sum_{k \in R_g} d_k q_k} \right)} \approx \frac{\sum_{g=1}^G \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k e_k}{\sum_{k \in R_g} d_k q_k} \right)}{\sum_{g=1}^G \hat{N}_g}, \quad (5)$$

where $e_k = y_k - \bar{y}_S$. Equation (5) holds exactly when all $q_k = 1$. When $q_k = p_k$, the near equality depends on the r_g being sufficiently large and other mild conditions.

Now assume the following response model holds: Each element k in a group has an equal, positive probability of response that does not vary with π_k or with y_k . That is to say, the response indicator ρ_k , which is 1 when k responds if sampled and is 0 otherwise, is a Bernoulli random variable with a common mean in U_g regardless of the values of π_k and y_k .

By treating unit response as a second phase of probability sampling in this way, the added variance/mean-squared-error due to nonresponse given the original sample and the r_g for both estimators can be expressed as

$$A_q = E_p[(\hat{y}_{S,q} - \bar{y}_S)^2 | S, \{r_g\}] \\ \approx \frac{\sum_{g=1}^G \hat{N}_g^2 \text{Var}_p(\hat{e}_{S,q} | S_g, r_g)}{\left(\sum_{g=1}^G \hat{N}_g \right)^2}, \quad (6)$$

where $\hat{e}_{S,q} = \hat{y}_{S,q} - \bar{y}_S$, $\bar{e}_{S_g} = \bar{y}_{S_g} - \bar{y}_S$, and

$$\begin{aligned} \text{Var}_p(\hat{e}_{S,q} | S_g, r_g) \\ \approx \left(\frac{n_g}{r_g} - 1 \right) \frac{\sum_{k \in S_g} d_k^2 (e_k - q_k \bar{e}_{S_g})^2}{\left(\sum_{k \in S_g} d_k q_k \right)^2} \\ = \left(\frac{n_g}{r_g} - 1 \right) \frac{\sum_{k \in S_g} d_k^2 ([y_k - \bar{y}_S] - q_k [\bar{y}_{S_g} - \bar{y}_S])^2}{\left(\sum_{k \in S_g} d_k q_k \right)^2}, \quad (7) \end{aligned}$$

under mild conditions on the population and original sampling design we assume to hold, including (again) that the r_g are sufficiently large. These conditions make both estimators nearly unbiased under quasi-probability sampling theory (probability theory augmented with a response model) and render the distinction between large-sample variance and mean squared error moot. Quasi-probability sampling theory is also known as “quasi-design-based” and “quasi-randomization-based” sampling theory.

Looking at equations (6) and (7), we see that at one extreme \hat{y}_{rw} has an added variance due to nonresponse of (approximately) zero when all the originally sampled y_k in a group are equal, while at the other \hat{y}_{de} has an added variance of zero when all the originally sampled $d_k e_k$ (or, put another way, the $d_k [y_k - \bar{y}_S]$) in a group are equal.

Heuristically, the reweighted-expansion estimator is more efficient than the double-expansion estimator when \bar{e}_{S_g} is a better predictor of e_k for $k \in S_g$ than $p_k \bar{e}_{S_g}$. Thus, when the groups were constructed as advised in Little and Vartivarian (2003) and earlier in Little (1986) so that the y_k in a group are homogeneous (as opposed to the $d_k [y_k - \bar{y}_S]$ being homogeneous), then the reweighted-expansion estimator computed with the f_g will usually be more efficient than the double-expansion estimator computed with the n_g/r_g .

The heuristic observation can be formalized with an alternative justification for using the reweighted-expansion estimator. Suppose the following prediction model holds: Each y_k in U_g is a random variable with common mean, μ_g , regardless of π_k and ρ_k . Then \hat{y}_{rw} is nearly unbiased under mild conditions with respect to the combination of the original sampling mechanism (which treats the d_k as random, where $d_k = 0$ for $k \neq S$) the prediction model (which treats the y_k as random). That is to say, $E_d[E_y(\hat{y}_{rw} - \bar{y}_U | S)] \approx 0$, since the double expectation of both \hat{y}_{rw} and \bar{y}_U are nearly $\sum_{g=1}^G N_g \mu_g / \sum_{g=1}^G N_g$. This combined unbiasedness is exact when the design is such that $\sum_S d_k = N$. Stratified, simple random sampling is an example of such a design. Unstratified sampling with unequal probabilities and many multistage designs are not.

It is not hard to see that \hat{y}_{rw} is also exactly unbiased with respect to this double expectation (i.e., $E_d[E_y(\hat{y}_{rw} - \bar{y}_U | S)] = 0$) when all the μ_g are equal. In fact, the prediction-model expectation of both \hat{y}_{rw} and \hat{y}_{de} equals this common mean, as does the prediction-model expectation of an estimator without any adjustment for unit nonresponse, that is, with the f_g in \hat{y}_{rw} replaced by 1. The advantage of \hat{y}_{rw} over \hat{y}_{de} under the prediction model obtains only when the μ_g vary, that is, when the survey variable has a different prediction mean across the groups.

Notice that if *either* the response model or the prediction model holds, then the reweighted-expansion estimator is nearly unbiased in some sense (i.e., under the combination of the original design and the response model or under the original design and the prediction model). This property has been called “double protection” against nonresponse bias. See, for example, Bang and Robins (2005).

3. Concluding remarks

In this note, we discussed two distinct types of models. We stressed a response model, which treats the response indicators, ρ_k , as a Bernoulli random variable within each group but with unknown parameters. We also described a prediction model, which treats the survey values, y_k , as random variables with unknown means that could vary across groups but not within them.

As part of the response model, we assumed that within a group, the ρ_k do not depend on the y_k . Analogously as part of the prediction model, we assumed that within a group, the y_k do not depend on the ρ_k . When both ρ_k and y_k are treated as random variables the former assumption, that nonrespondents are *missing at random*, is equivalent to the latter assumption, that the response mechanism is *ignorable* (see, for example, Little and Rubin 1987). It should be understood, however, that the y_k need not be treated as random variables under the response model and the ρ_k need not be treated as random variables under the prediction model. The two concepts (missingness at random and ignorable non-response) may be equivalent in some sense but they are not identical.

The heart of Little and Vartivarian (2003) is a series of simulations featuring a binary survey variable, two potential response groups, and two original selection probabilities. Both the survey variable and response indicators are generated under five models. The expected value of each is a function of, 1, the response group only, 2, the selection probability only, 3, neither, or, 4 and 5, one of two equal combinations of response group and selection probability. This produces 25 scenarios, of which 10 are of primary interest to us. Those are the ones in which the survey

variable is a function either of only the response group or of neither the response group nor the selection probability.

As our theory predicts when the survey variable is a function of neither the response group nor the selection probability, both the reweighted- and double-expansion estimators have empirical biases near zero (Table 5 in Little and Vartivarian) because both are nearly unbiased under the combination of the original sampling design and a valid prediction model: all population elements have the same mean. When the survey variable is a function of the response group *and* the response indicator is wholly or partly a function of the selection probability, only the reweighted-expansion estimator is nearly unbiased empirically since only it is unbiased under the combination of the original sampling design and a valid prediction model. As a result, \hat{y}_{rw} also has less empirical root mean squared error and significantly less average absolute error as an estimator for \bar{y}_S (Tables 4 and 6 in Little and Vartivarian, respectively; the significance test treats the mean value across the simulations of $|\hat{y}_{rw} - \bar{y}_S| - |\hat{y}_{de} - \bar{y}_S|$ as asymptotically normal).

When both the survey variable and response indicators are functions of the response group only, the reweighted-expansion estimator has slightly less empirical root mean squared error and average absolute error than the double-expansion estimator but the latter is not significant.

It should not surprise us that the reduction in empirical root mean squared error is modest. The contribution to the variance from nonresponse under the response model mechanism expressed in equations (6) and (7) is conditioned on the original sample (technically, the contribution of non-response to the total quasi-probability variance of $\hat{y}_{S,q}$ is the expectation of A_q in equation (6) under the original sampling mechanism). In applications where the response rates are relatively large (in the simulations they averaged 0.5), this contribution can be dominated by the probability-sampling variance/mean squared error of the full-sample estimator, \hat{y}_U .

Two warnings are in order. The respondent sample size within each group must be sufficiently large for the reweighted-expansion estimator to nearly unbiased under quasi-probability theory. For the double-expansion estimator, each r_g need only be positive. Moreover, that the reweighted-expansion estimator is doubly protected against nonresponse bias is only helpful when either the assumed response or prediction model is correct. If *both* the response probabilities and survey values vary with the design weights, then the reweighted-expansion estimator can be meaningfully biased. Despite the slant taken in this note, that is the take-away message Little and Vartivarian (2003) intended, and it cannot be disputed.

Acknowledgements

I thank the associate editor and two referees whose careful reading of earlier versions of the manuscript greatly improved the quality and accuracy of the resulting work. Any remaining errors in the text are wholly due to the author.

References

- Bang, H., and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-972.
- Fuller, W. (2009). *Sampling Statistics*, Hoboken, New Jersey: Wiley.
- Kim, J.K., Navarro, A. and Fuller, W. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Little, R. (1986). Survey nonresponse adjustments. *International Statistical Review*, 54, 139-157.
- Little, R., and Rubin, D. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.
- Little, R., and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.
- Lohr, S (2009). *Sampling: Design and Analysis, Second Edition*, Boston: Brooks/Cole.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 27, No. 4, 2011

An Index Number Formula Problem: The Aggregation of Broadly Comparable Items Mick Silver.....	553
Identifying Sources of Error in Cross-national Questionnaires: Application of an Error Source Typology to Cognitive Interview Data Rory Fitzgerald, Sally Widdop, Michelle Gray, Debbie Collins.....	569
Using a Geographic Segmentation to Understand, Predict, and Plan for Census and Survey Mail Nonresponse Nancy Bates, Mary H. Mulry	601
Are They Really Too Busy for Survey Participation? The Evolution of Busyness and Busyness Claims in Flanders Anina Vercruyssen, Bart van de Putte, Ineke A.L. Stoop	619
Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys Avinash C. Singh, Fulvia Mecatti	633
On Variance Estimation for the French Master Sample Guillaume Chauvet	651
Secure Multiple Linear Regression Based on Homomorphic Encryption Rob Hall, Stephen E. Fienberg, Yuval Nardi.....	669
Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey Nicole Watson, Rosslyn Starick	693
Book Reviews.....	717
Editorial Collaborators	723
Index to Volume 27, 2011	727

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 28, No. 1, 2012

Capture–Recapture Sampling and Indirect Sampling Pierre Lavallée, Louis-Paul Rivest	1
An Examination of Within-Person Variation in Response Propensity over the Data Collection Field Period Kristen Olson, Robert M. Groves	29
Concording U.S. Harmonized System Codes over Time Justin R. Pierce, Peter K. Schott	53
Uncertainty Analysis in Statistical Matching Pier Luigi Conti, Daniela Marella, Mauro Scanu	69
Constrained Small Area Estimators Based on M-quantile Methods Enrico Fabrizi, Nicola Salvati, Monica Pratesi	89
A Multiplicative Masking Method for Preserving the Skewness of the Original Micro-records Nicolas Ruiz	107
Cell Bounds in k -way Tables Given Conditional Frequencies Byran J. Smucker, Aleksandra Slavković, Xiaotian Zhu	121
Book Review	141
In Other Journals	151

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Volume 39, No. 4, December/décembre 2011

Zhiguo Li and Bin Nan	
Relative risk regression for current status data in case-cohort studies	557
Liuquan Sun, Xingqiu Zhao and Jie Zhou	
A class of mixed models for recurrent event data	578
Chong Gu and Ping Ma	
Nonparametric regression with cross-classified responses	591
David R. Bickel	
A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons.....	610
Xiaoming Wang and Keumhee C. Carriere	
Assessing additivity in nonparametric models - A kernel-based method.....	632
Chenlei Leng and Cheng Yong Tang	
Improving variance function estimation in semiparametric longitudinal data analysis	656
Jihnhee Yu, Albert Vexler, Seong-Eun Kim and Alan D. Hutson	
Two-sample empirical likelihood ratio tests for medians in application to biomarker evaluations	671
Marwan Zidan, Jung-Chao Wang and Magdalena Niewiadomska-bugaj	
Comparison of k independent, zero-heavy lognormal distributions.....	690
Ivan Kojadinovic, Johan Segers and Jun Yan	
Large-sample tests of extreme-value dependence for multivariate copulas	703
Jeffrey S. Rosenthal	
Was the conservative majority predictable?	721
Errata	
Hideatsu Tsukahara	
Erratum to "Semiparametric estimation in copula models"	734

CONTENTS

TABLE DES MATIÈRES

Volume 40, No. 1, March/mars 2012

D.M. Hajducek and J.F. Lawless Duration analysis in longitudinal studies with intermittent observation times and losses to followup	1
Yongseok Park, John D. Kalbfleisch and Jeremy M.G. Taylor Constrained nonparametric maximum likelihood estimation of stochastically ordered survivor functions.....	22
Binbing Yu and Ram C. Tiwari A Bayesian approach to mixture cure models with spatial frailties for population-based cancer relative survival data	40
Yasaman Hosseinkashi, Shojaeddin Chenouri, Christopher G. Small and Rob Deardon A stochastic graph process for epidemic modelling.....	55
E.C. Brechmann, C. Czado and K. Aas Truncated regular vines in high dimensions with application to financial data	68
Alexander Bauer, Claudia Czado and Thomas Klein Pair-copula constructions for non-Gaussian DAG models.....	86
Liang Peng Approximate jackknife empirical likelihood method for estimating equations	110
Guillaume Chauvet and David Haziza Fully efficient estimation of coefficients of correlation in the presence of imputed survey data.....	124
Zilin Wang and Mary E. Thompson A resampling approach to estimate variance components of multilevel models.....	150
Huybrechts F. Bindele and Asheber Abebe Bounded influence nonlinear signed-rank regression	172
Mary C. Meyer Constrained penalized splines	190

GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

1. Layout

- 1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The documents should be divided into numbered sections with suitable verbal titles.
- 1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de finaliser votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1. **Présentation**
 - 1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0 ; l, I).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4. **Figures et tableaux**

Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6. **Communications brèves**

Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.

TABLE DES MATIÈRES

CONTENTS

Volume 40, No. 1, March/mars 2012

D.M. Hájduček and J.F. Lawless	1
Duration analysis in longitudinal studies with intermittent observation times and losses to followup	1
Yongseok Park, John D. Kalbfleisch and Jeremy M.G. Taylor	22
Constrained nonparametric maximum likelihood estimation of stochastically ordered survivor functions.....	22
Binbing Yu and Ram C. Tiwari	40
A Bayesian approach to mixture cure models with spatial frailties for population-based cancer relative survival data	40
Yasaman Hosseinkashi, Shojaeddin Chenouri, Christopher G. Small and Rob Deardon	55
A stochastic graph process for epidemic modelling.....	55
E.C. Brechmann, C. Czado and K. Aas	68
Truncated regular vines in high dimensions with application to financial data	68
Alexander Bauer, Claudia Czado and Thomas Klein	86
Pair-copula constructions for non-Gaussian DAG models.....	86
Liang Peng	110
Approximate jackknife empirical likelihood method for estimating equations	110
Guillaume Chauvet and David Haziza	124
Fully efficient estimation of coefficients of correlation in the presence of imputed survey data.....	124
Zilin Wang and Mary E. Thompson	150
A resampling approach to estimate variance components of multilevel models.....	150
Huybrechts F. Bindele and Ashbeher Abebe	172
Bounded influence nonlinear rank regression	172
Mary C. Meyer	190
Constrained penalized splines	190

Volume 39, No. 4, December/décembre 2011

Zhiguo Li and Bin Nan	Relative risk regression for current status data in case-cohort studies	557
Liuquan Sun, Xingqiu Zhao and Jie Zhou	A class of mixed models for recurrent event data	578
Chong Gu and Ping Ma	Nonparametric regression with cross-classified responses	591
David R. Bickel	A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons	610
Xiaoming Wang and Keumhee C. Carrière	Assessing additivity in nonparametric models - A kernel-based method	632
Chenlei Leng and Cheng Yong Tang	Improving variance function estimation in semiparametric longitudinal data analysis	656
Jihhee Yu, Albert Vexler, Seong-Eun Kim and Alan D. Hutson	Two-sample empirical likelihood ratio tests for medians in application to biomarker evaluations	671
Marwan Zidan, Jung-Chao Wang and Magdalena Niewiadomska-bugaj	Comparison of k independent, zero-heavy lognormal distributions	690
Ivan Kojadinovic, Johan Segers and Jun Yan	Large-sample tests of extreme-value dependence for multivariate copulas	703
Jeffrey S. Rosenthal	Was the conservative majority predictable?	721
Hideatsugu Tsukahara	Erratum to "Semiparametric estimation in copula models"	734

Errata

JOURNAL OF OFFICIAL STATISTICS
An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 28, No. 1, 2012

Capture–Recapture Sampling and Indirect Sampling	1
Pierre Lavallée, Louis-Paul Rivest	
An Examination of Within-Person Variation in Response Propensity over the Data Collection Field Period	29
Kristen Olson, Robert M. Groves	
Concording U.S. Harmonized System Codes over Time	53
Justin R. Pierce, Peter K. Schott	
Uncertainty Analysis in Statistical Matching	69
Pier Luigi Coniti, Daniela Marella, Mauro Scanu	
Constrained Small Area Estimators Based on M-quantile Methods	89
Enrico Fabrizi, Nicola Salvati, Monica Pratesi	
A Multiplicative Masking Method for Preserving the Skewness of the Original Micro-records	107
Nicolas Ruiz	
Cell Bounds in <i>k</i> -way Tables Given Conditional Frequencies	121
Byran J. Smucker, Aleksandra Slavković, Xiaotian Zhu	
Book Review	141
In Other Journals	151

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 27, No. 4, 2011

An Index Number Formula Problem: The Aggregation of Broadly Comparable Items	Mick Silver	553
Identifying Sources of Error in Cross-national Questionnaires: Application of an Error Source Typology	Rory Fitzgerald, Sally Widdop, Michelle Gray, Debbie Collins	569
Using a Geographic Segmentation to Understand, Predict, and Plan for Census and Survey Mail Nonresponse	Nancy Bates, Mary H. Muly	601
Are They Really Too Busy for Survey Participation? The Evolution of Business and Business Claims in Flanders	Anina Vercruyssen, Bart van de Putte, Ineke A.L. Stoop	619
Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys	Avinash C. Singh, Fulvia Mecatti	633
On Variance Estimation for the French Master Sample	Guillaume Chauvet	651
Secure Multiple Linear Regression Based on Homomorphic Encryption	Rob Hall, Stephen E. Fienberg, Yuval Nardi	669
Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey	Nicole Watson, Rosslyn Starck	693
Book Reviews		717
Editorial Collaborators		723
Index to Volume 27, 2011		727

All inquiries about submissions and subscriptions should be directed to jos@scb.se

non-réponse n'est utile que si le modèle de réponse ou le modèle de prédiction hypothétique est correct. Si les probabilités de réponse *ainsi que* les valeurs des variables observées varient avec les poids de sondage, l'estimateur à facteur d'extension répondre peut présenter un biais significatif. Malgré la perspective adoptée dans le présent exposé, il s'agit du message que Little et Vartivarian (2003) souhaitaient communiquer et il ne peut pas être contesté.

Remerciements

Je remercie le rédacteur associé et deux examinateurs de leur lecture attentive de versions antérieures du manuscrit qui m'a permis d'améliorer considérablement la qualité et l'exactitude des travaux résultants. Toute erreur qui subsiste dans le texte est entièrement imputable à l'auteur.

Bibliographie

- Bang, H., et Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-972.
- Fuller, W. (2009). *Sampling Statistics*. Hoboken, New Jersey : Wiley.
- Kim, J.K., Navarro, A. et Fuller, W. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Little, R. (1986). Survey nonresponse adjustments. *Revue Internationale de Statistique*, 54, 139-157.
- Little, R., et Rubin, D. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley & Sons, Inc.
- Little, R., et Vartivarian, S. (2003). On weighing the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.
- Lohr, S. (2009). *Sampling: Design and Analysis, Second Edition*. Boston : Brooks/Cole.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

facteur d'extension répondre est presque sans biais empiriquement, puisqu'il est le seul à être sans biais sous la combinaison du plan d'échantillonnage original et d'un modèle de prédiction valide. Par conséquent, \hat{y}_{rw} donne aussi une racine carrée de l'erreur quadratique moyenne empirique plus faible et une erreur absolue moyenne significativement plus faible qu'un estimateur de \bar{y}_s (tableaux 4 et 6 dans Little et Vartivarian, respectivement ; le test de signification traite la valeur moyenne sur l'ensemble des simulations de $|\hat{y}_{rw} - \bar{y}_s| - |\hat{y}_{de} - \bar{y}_s|$ comme étant asymptotiquement normale).

Quand la variable d'enquête ainsi que les indicateurs de réponse sont des fonctions du groupe de réponse seulement, l'estimateur à facteur d'extension répondre produit une racine carrée de l'erreur quadratique moyenne empirique et une erreur absolue moyenne légèrement plus faibles que l'estimateur à facteur d'extension double, mais la différence n'est pas significative pour le second paramètre.

Il n'est guère surprenant que la réduction de la racine carrée de l'erreur quadratique moyenne empirique soit modeste. La contribution de la non-réponse à la variance sous le modèle de réponse exprimé par les équations (6) et (7) est conditionnelle à l'échantillon original (techniquement, la contribution de la non-réponse à la variance totale sous échantillonnage quasi probabiliste de \hat{y}_{sq} est l'espace de \mathcal{A}_q dans l'équation (6) sous le mécanisme d'échantillonnage original). Dans les applications où les taux de réponse sont assez élevés (dans les simulations ils valent en moyenne 0,5), cette contribution peut être dominée par la variance/l'erreur quadratique moyenne sous échantillonnage probabiliste de l'estimateur sous échantillon complet, \hat{y}_{Uq} .

Deux mises en garde sont de rigueur. La taille de l'échantillon de répondants dans chaque groupe doit être suffisamment grande pour que l'estimateur à facteur d'extension répondre soit presque sans biais sous la théorie de l'échantillonnage quasi probabiliste. Pour l'estimateur à facteur d'extension double, il suffit que chaque taille r_s soit positive. En outre, le fait que l'estimateur à facteur d'extension répondre est doublement protégé contre le biais de

3. Conclusion

Le présent exposé porte sur deux types de modèles distincts. Nous avons décrit un modèle de réponse dans lequel les indicateurs de réponse, p_k , sont traités comme une variable aléatoire de Bernoulli dans chaque groupe, mais dont les paramètres sont inconnus. Nous avons également décrit un modèle de prédiction dans lequel les valeurs observées, y_k , sont traitées comme des variables aléatoires de moyenne inconnue pouvant varier entre les groupes mais non à l'intérieur de ceux-ci.

Dans le modèle de réponse, nous supposons qu'à l'intérieur d'un groupe, les p_k ne dépendent pas des y_k . Par analogie, dans le modèle de prédiction, nous supposons que, dans un groupe, les y_k ne dépendent pas des p_k . Quand p_k ainsi que y_k sont traités comme des variables aléatoires, la première hypothèse, à savoir que les non-répondants *manquent au hasard*, équivaut à la seconde hypothèse, à savoir que le mécanisme de réponse est *ignorable* (voir, par exemple, Little et Rubin 1987). Il convient toutefois de bien saisir que les y_k n'ont pas à être traités comme des variables aléatoires sous le modèle de réponse et que les p_k n'ont pas à être traités comme des variables aléatoires sous le modèle de prédiction. Les deux concepts (réponse manquant au hasard et non-réponse ignorable) sont peut-être équivalents dans un certain sens, mais ils ne sont pas identiques.

L'exposé de Little et Vartivarian (2003) a pour élément central une série de simulations comportant une variable d'enquête binaire, deux groupes de réponses possibles et deux probabilités de sélection originales. La variable d'enquête ainsi que les indicateurs de réponse sont générés sous cinq modèles. La valeur prévue de chacune de ces variables (1) dépend du groupe de réponse seulement, 2) dépend de la probabilité de sélection seulement, 3) ne dépend ni de l'un ni de l'autre, ou 4) et 5) dépend de l'une de deux combinaisons égales de groupe de réponse et de probabilité de sélection. Cela produit 25 scénarios dont dix nous intéressent tout spécialement. Ces dix scénarios sont ceux dans lesquels la variable d'enquête est une fonction du groupe de réponse seulement ou n'est une fonction ni du groupe de réponse ni de la probabilité de sélection.

Comme le prédit notre théorie quand la variable d'enquête n'est fonction ni du groupe de réponse ni de la probabilité de sélection, l'estimateur à facteur d'extension répondra à deux biais empiriques presque nuls (tableau 5 dans Little et Vartivarian) parce qu'ils sont tous deux presque sans biais d'un modèle de prédiction valide : toutes les unités de la population ont la même moyenne. Quand la variable d'enquête est une fonction du groupe de réponse et que l'indicateur de réponse est entièrement ou partiellement une fonction de la probabilité de sélection, seul l'estimateur à

$d_k e_k$ (ou, exprimés d'une autre façon, les $d_k [y_k - \bar{y}_S]$ échantillonnés au départ dans un groupe sont égaux.

Sur le plan heuristique, l'estimateur à facteur d'extension répondra est plus efficace que l'estimateur à facteur d'extension double quand \bar{e}_S^2 est un meilleur prédicteur de e_k que $p_k \bar{e}_S$ pour $k \in S_g$. Donc, quand les groupes sont constitués comme l'ont recommandé Little et Vartivarian (2003), et apparemment Little (1986), de façon à ce que la réponse y_k dans un groupe soit homogène (par opposition au fait que $d_k [y_k - \bar{y}_S]$ soit homogène), l'estimateur à facteur d'extension répondra sera plus efficace que l'estimateur à facteur d'extension double calculé en se servant de f_g sera généralement plus efficace que l'estimateur à facteur d'extension double calculé en se servant de n_g^2/r_g^2 .

L'observation heuristique peut être exprimée formellement en utilisant une autre justification de l'estimateur à facteur d'extension répondra. Supposons que le modèle de prédiction qui suit est vérifié : Chaque y_k dans U_g est une variable aléatoire de moyenne commune, μ_g , indépendamment de π_k et p_k . Alors, \bar{y}_{ng} est presque sans biais sous des contraintes faibles en ce qui concerne la combinaison du mécanisme d'échantillonnage original (qui traite les d_k comme étant aléatoires, où $d_k = 0$ pour $k \neq S$) et du modèle de prédiction (qui traite les y_k comme étant aléatoires). C'est-à-dire que $E_y[E_y(\bar{y}_{ng}) - \bar{y}_U | S]] \approx 0$, puisque l'espérance double de \bar{y}_{ng} ainsi que \bar{y}_U est presque $\Sigma p_k \mu_g / \Sigma p_k$. Cette absence combinée de biais est exacte quand le plan est tel que $\Sigma_S d_k \equiv N$. L'échantillonnage aléatoire simple stratifié est un exemple de ce type de plan. L'échantillonnage non stratifié avec probabilités inégales et de nombreux plans à plusieurs degrés ne le sont pas.

Il n'est pas difficile de voir que \bar{y}_{ng} est également sans biais par rapport à cette espérance double (c'est-à-dire $E_y[E_y(\bar{y}_{ng}) - \bar{y}_U | S]] = 0$) quand tous les μ_g sont égaux. En fait, l'espérance de \bar{y}_{ng} ainsi que de \bar{y}_{de} sous le modèle de prédiction est égale à cette moyenne commune, de même que l'espérance sous le modèle de prédiction d'un estimateur sans aucune correction pour tenir compte de la non-réponse totale, c'est-à-dire avec le remplacement de f_g dans \bar{y}_{ng} par 1. L'avantage de \bar{y}_{ng} par rapport à \bar{y}_{de} sous le modèle de prédiction s'obtient uniquement quand les μ_g varient, c'est-à-dire quand la moyenne de prédiction de la variable d'enquête varie d'un groupe à l'autre.

Il convient de souligner que si le modèle de réponse ou le modèle de prédiction tient, l'estimateur à facteur d'extension répondra est presque sans biais dans un certain sens (c'est-à-dire sous la combinaison du plan de sondage original et du modèle de réponse ou sous le plan de sondage appelé « double protection » contre le biais de non-réponse. Voir, par exemple, Bang et Robins (2005).

L'emploi de n_k/r_k donne l'estimateur à facteur d'extension double :

$$\hat{y}_{S,q}^{de} = \frac{\sum_{g=1}^G \left(\frac{1}{n_k} \sum_{k \in R_k^g} d_k y_k \right)}{\sum_{g=1}^G \left(\frac{1}{n_k} \sum_{k \in R_k^g} d_k \right)}.$$

Pour les besoins de notre étude, cet estimateur peut également être exprimé sous la forme

$$\hat{y}_{S,q}^{de} = \frac{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)}{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)} = \frac{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)}{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)}.$$

où

$$(3) \quad d_k = \frac{1}{\sum_{j \in S_k^g} d_j} \quad \text{pour } k \in S_k^g$$

(de sorte que $\sum_{k \in S_k^g} d_k \bar{d}_k = \sum_{k \in S_k^g} d_k = N_k^g$). Les estimateurs $\hat{y}_{S,q}^{nu}$ et $\hat{y}_{S,q}^{de}$ peuvent s'écrire tous deux sous la forme :

$$(4) \quad \hat{y}_{S,q}^{nu} = \frac{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)}{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)}.$$

Pour l'estimateur à facteur d'extension pondéré, tous les $q_k = 1$, tandis que pour l'estimateur à facteur d'extension double, $q_k = d_k$ tel qu'il est défini par l'équation (3). Nous nous servons bientôt de l'expression qui suit pour nos deux estimateurs :

$$(5) \quad \hat{y}_{S,q}^{de} - \bar{y}_S = \frac{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)}{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)} - \bar{y}_S = \frac{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)}{\left(\sum_{g=1}^G \frac{\sum_{k \in R_k^g} d_k y_k}{\sum_{k \in R_k^g} d_k} \right)} - \bar{y}_S.$$

où $e_k = y_k - \bar{y}_S$. L'équation (5) est vérifiée exactement quand tout $q_k = 1$. Si $q_k = d_k$, la presque égalité

dépendra du fait que r_k est suffisamment grand, ainsi que d'autres contraintes faibles. Supposons maintenant que le modèle de réponse qui suit est vérifié : Chaque unité k d'un groupe a une probabilité positive de réponse qui ne varie pas en fonction de π_k ni de y_k . Autrement dit, l'indicateur de réponse p_k qui vaut 1 quand l'unité k répond si elle est échantillonnée et 0 autrement, est une variable aléatoire de Bernoulli dont la moyenne est commune dans U_k indépendamment des valeurs de π_k et y_k . En traitant de cette façon la réponse d'une unité comme une deuxième phase de l'échantillonnage probabiliste, nous pouvons exprimer la variance/l'erreur quadratique moyenne supplémentaire due à la non-réponse, sachant l'échantillon original et les r_k pour les deux estimateurs, sous la forme

$$(6) \quad A_q = E_p(\hat{y}_{S,q}^2 | S, \{r_k\}) \approx \frac{\sum_{g=1}^G N_k^g \text{Var}_p(\hat{e}_{S,q}^2 | S_k^g, r_k^g)}{\left(\sum_{g=1}^G N_k^g \right)^2},$$

où $\hat{e}_{S,q}^2 = \hat{y}_{S,q}^2 - \bar{y}_S^2$, $\bar{e}_{S,q}^2 = \bar{y}_S^2 - \bar{y}_S^2$, et

$\text{Var}_p(\hat{e}_{S,q}^2 | S_k^g, r_k^g)$

$$(7) \quad = \left(\frac{n_k^g}{n_k^g} - 1 \right) \frac{\left(\sum_{k \in S_k^g} d_k y_k \right)^2}{\sum_{k \in S_k^g} d_k^2 (e_k - q_k \bar{e}_{S,q}^2)} - \left(\frac{n_k^g}{n_k^g} - 1 \right) \frac{\left(\sum_{k \in S_k^g} d_k y_k \right)^2}{\sum_{k \in S_k^g} d_k^2 (e_k - q_k \bar{e}_{S,q}^2)}.$$

en appliquant à la population et au plan de sondage original des contraintes faibles que nous supposons être vérifiées, y des contraintes faibles (de nouveau) le fait que les tailles r_k sont suffisamment grandes. Ces conditions rendent les deux estimateurs presque sans biais sous la théorie de l'échantillonnage quasi probabiliste (théorie probabiliste augmentée d'un modèle de réponse) et rend discutable la distinction entre la variance et l'erreur quadratique moyenne en grand échantillon. La théorie de l'échantillonnage quasi probabiliste est également appelée théorie de l'échantillonnage « quasi fondée sur un plan » ou « quasi aléatoire ».

Si nous examinons les équations (6) et (7), nous constatons qu'à l'un des extrêmes, $\hat{y}_{S,q}^{nu}$ possède une variance supplémentaire due à la non-réponse (approximativement) nulle quand tous les y_k échantillonnés originellement dans un groupe sont égaux, tandis qu'à l'autre extrême, $\hat{y}_{S,q}^{de}$ présente une variance supplémentaire nulle quand tous les

Supposons que nous voulions estimer la moyenne de population d'une variable d'enquête y_k :

2. Les deux estimateurs

pas. À la section 3, nous montrons que les résultats empiriques présentés dans Little et Vartivarian (2003) concordent avec ces arguments et nous offrons certaines conclusions.

A la section 2, nous établissons la notation pour estimer la moyenne de population d'une variable d'enquête, l'utilisation du rapport n_g/r_g produit un estimateur à facteur d'extension double tandis que l'utilisation de f_g^s produit un estimateur à facteur d'expansion pondérée. Nous pouvons exprimer les deux estimateurs en nous servant d'une formulation donnée dans Kim, Navarro et Fuller (2006). Cette expression permet de voir que, si la variable d'enquête se comporte approximativement comme une variable aléatoire de moyenne constante à l'intérieur de chaque groupe, quels que soient les poids de sondage, l'utilisation de f_g^s est souvent plus efficace que l'utilisation de n_g/r_g . En fait, si la variable d'enquête se comporte exactement comme une telle variable aléatoire, l'estimation de la moyenne de population calculée en se servant de f_g^s sera presque sans biais sous la combinaison du plan de sondage original et de ce modèle de prédiction, même si le modèle de réponse ne tient

Little et Varianian (2003) affirment que f^g est habituellement utilisé en pratique. Toutefois, ils soulignent qu'intégrer de cette façon les poids de sondage dans le facteur de correction peut « accroître la variance ».

où S_g est l'échantillon original et R_g le sous-échantillon de répondants dans le groupe g . Ce facteur de correction peut différer de n_g/r_g quand les d_k dans le groupe g

$$(I) \quad \sum_{k \in S} p_k \sum_{k \in R} p_k = f$$

Bien que la probabilité conditionnelle de réponse dans le groupe g sous le modèle de réponse bernaoullien stratifié soit d^g/n^g , nous verrons qu'il est souvent préférable de multiplier le poids de sondage, $d^g = 1/\pi^g$, d'une unité répondante dans le groupe non pas par d^g/n^g , mais par

Conditionnellement, aux tailles des échantillons de répondants dans les groupes, un sous-échantillon benjaminien stratifié dont les probabilités de sélection (réponses) sont inconnues est converti en un sous-échantillon aléatoire simple stratifié dont les probabilités de sélection sont connues : π_{1h}^* / π_{2h}^* mod les unités d'un groupe g quand ce groupe contient u unités

nologie plus simple ici.

Techniquement, $\frac{Y}{Y^w}$ est le ratio de deux estimateurs à facteur d'extension répondère, mais nous utilisons la terminologie plus simple ici.

[illegible]

Si nous faisons une correction pour tenir compte de la non-réponse en utilisant le facteur f^{δ} dans l'équation (1), nous obtenons l'estimateur à facteur d'extension pondéré :

Fuller (2006).
 Désignons l'estimateur de \hat{y}_U sous échantillon complet
 dont nous avons discuté par $\hat{y}_S = \sum_{i \in N} \hat{y}_S^i$. Il existe des
 moyens plus directs de rendre \hat{y}_S , mais la version susmen-

grand. Nous émettons ces deux hypothèses ici.

sous des contraintes faibles quand n^g est suffisamment

$$(2) \quad \frac{\sum_{k \in S^*} p_k}{\sum_{k \in S} p_k} = \frac{p_{S^*}}{p_S}$$

[illegible]

$$\frac{\sum_{\mathcal{O}}^8 N \sum_{\mathcal{O}}^{l=8}}{\sum_{\mathcal{O}}^8 N \sum_{\mathcal{O}}^{l=8}} = \frac{\sum_{\mathcal{O}}^8 N \sum_{\mathcal{O}}^{l=8}}{\sum_{\mathcal{O}}^8 N \sum_{\mathcal{O}}^{l=8}} = \frac{N}{\sum_{\mathcal{O}}^8} = \underline{n_{\mathcal{A}}}$$

Kott : Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale

Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes

Philip S. Kott¹

Résumé

En cas de non-réponse totale d'une unité dans un échantillon tiré suivant les principes de l'échantillonnage probabiliste, une pratique courante consiste à diviser l'échantillon en groupes mutuellement exclusifs de manière qu'il soit raisonnable de supposer que toutes les unités échantillonnées dans un groupe ont la même probabilité de ne pas répondre. De cette façon, la réponse d'une unité peut être traitée comme une phase supplémentaire de l'échantillonnage probabiliste en se servant de l'inverse de la probabilité de réponse estimée d'une unité dans un groupe comme facteur de correction pour calculer les poids finaux pour les répondants du groupe. Si l'objectif est d'estimer la moyenne de population d'une variable d'enquête qui se comporte plus ou moins comme une variable aléatoire dont la moyenne est constante dans chaque groupe indépendamment des poids de sondage originaux, il est habituellement plus efficace d'intégrer les poids de sondage dans les facteurs de correction que de ne pas le faire. En fait, si la variable d'enquête se comportait exactement comme une telle variable aléatoire, l'estimation de la moyenne de population calculée en se servant des facteurs de correction pondérés selon le plan de sondage serait presque sans biais dans un certain sens (c'est-à-dire sous la combinaison du mécanisme d'échantillonnage probabiliste original et d'un modèle de prédiction), même si les unités échantillonnées dans un groupe n'ont pas toutes la même probabilité de répondre.

Mots clés : Double protection ; modèle de prédiction ; échantillonnage probabiliste ; modèle de réponse ; phase d'échantillonnage ; échantillonnage bernoullien stratifié.

1. Introduction

En l'absence de non-réponse, il est possible d'estimer la moyenne d'une population fine d'après un échantillon sans avoir à recourir à un modèle statistique qui, aussi raisonnable qu'il soit, pourrait ne pas être vérifié. Pour cela, on attribue à chaque unité de la population une probabilité positive de sélection dans l'échantillon et l'on crée des estimateurs en s'appuyant sur ce mécanisme de sélection aléatoire. Malheureusement, dans des conditions réelles, les enquêtes souffrent souvent de non-réponse.

Deux types distincts de modèle peuvent être utilisés pour faire face à la non-réponse totale d'une unité. L'un est un modèle de prédiction, ou de résultat, dans lequel on suppose que la variable d'enquête se comporte comme une variable aléatoire dont on connaît les caractéristiques, mais non les paramètres. L'autre est un modèle de réponse, ou de sélection, dans lequel le simple fait qu'une unité réponde à une enquête est traité comme une phase supplémentaire de la

sélection aléatoire de l'échantillon.

Habituellement, les statisticiens d'enquête préfèrent les modèles de réponse pour deux raisons. Outre le fait que la

modélisation de la réponse est comme de la magie, elle permet de traiter la réponse d'une unité comme une phase supplémentaire de l'échantillonnage aléatoire, une enquête est habituellement conçue afin de recueillir des renseignements sur plusieurs variables auprès des unités échantillonnées. La modélisation de la prédiction requiert que l'on formule un modèle hypothétique différent pour chaque

variable d'enquête, chacun de ces modèles pouvant ne pas être vérifié. Par contre, la modélisation de la réponse ne nécessite l'hypothèse que d'un seul modèle. Il n'en est toutefois plus ainsi en cas de non-réponse partielle (pour une variable particulière de l'enquête). Par conséquent, les modèles de prédiction sont souvent préférés pour traiter la non-réponse partielle par imputation. Cela étant dit, la non-réponse partielle dépasse le cadre du présent article.

Sous un modèle de réponse hypothétique, les probabilités de réponse des unités sont traitées comme étant inconnues, ce qui signifie qu'elles doivent être estimées d'après l'échantillon. Habituellement, on suppose que le mécanisme de réponse des diverses unités est indépendant et qu'il ne dépend pas de la sélection de l'unité dans l'échantillon (chaque unité possède une probabilité a priori de réponse qui devient opérationnelle si elle est sélectionnée dans l'échantillon). Le modèle de réponse le plus simple et le plus fréquemment utilisé consiste à diviser l'échantillon et, implicitement la population complète, en groupes mutuellement exclusifs, appelés « groupes de réponse homogènes » par Särndal, Swensson et Wretman (1992) (le terme « classes de pondération » est plus courant ; voir, par exemple, Lohr [2009, pages 340-341]), et à supposer que chaque unité d'un groupe possède la même probabilité de ne pas répondre, quelle que soit sa probabilité de sélection dans l'échantillon original, π_i . Donc, le mécanisme de réponse produit un sous-échantillon bernoullien stratifié dans lequel les groupes constituent les strates.

Annexe B

Modèles de non-réponse ignorable et non ignorable

Posons que $\psi^{ijk} \equiv 1$ dans le modèle d'extension pour former le modèle de non-réponse ignorable. Pour $i = 1, \dots, A_i$, nous prenons alors

$$\pi_i | \mu_2, \tau_2 \sim \text{Dirichlet}(\mu_2 \tau_2)$$

et indépendamment

$$p_i | \mu_1, \tau_1 \sim \text{Dirichlet}(\mu_1 \tau_1)$$

En outre, $p(\tau_2) = \{1/(1 + \tau_2^2)\}^2, \tau_1 \geq 0, \mu_1 \sim \text{Dirichlet}(1)$, $p(\tau_2) = \{1/(1 + \tau_2^2)\}^2, \tau_1 \geq 0$ et $\mu_2 \sim \text{Dirichlet}(1)$. Ici, nous avons l'indépendance à tous les niveaux et les vecteurs $\mathbf{1}$ sont de dimension appropriée, chaque coordonnée étant égale à l'unité. Notons que tous les paramètres du modèle ignorable doivent être identifiés et estimés.

Soit $\pi^{ijk} = \pi_i \psi^{ijk}$ dans le modèle d'extension pour former le modèle de non-réponse non ignorable. Dans ce cas, pour $i = 1, \dots, A_i$

$$\pi^{ijk} | \mu_2, \tau_2 \sim \text{Dirichlet}(\mu_2 \tau_2)$$

et indépendamment

$$p_i | \mu_1, \tau_1 \sim \text{Dirichlet}(\mu_1 \tau_1)$$

Dans ce modèle, les paramètres π^{ijk} ne sont pas identifiables et nous prenons $\tau_2 \sim \text{Gamma}(\alpha_0, \beta_0)$, où α_0 et β_0 doivent être spécifiés. La spécification du modèle est alors achevée en attribuant à τ_1, μ_1 et à μ_2 les mêmes propriétés distributionnelles qu'au paragraphe précédent.

Comme dans Nandram, Cox et Choi (2005), α_0 et β_0 sont spécifiés comme il suit. Le modèle de non-réponse ignorable et ajusté de manière à obtenir un échantillon de la densité de probabilité à posteriori de τ_2 . Puis, α_0 et β_0 sont obtenus en utilisant la méthode des moments. Nandram, Cox et Choi (2005) ont constaté que l'inférence au sujet de p_i n'est pas très sensible au choix de ces paramètres.

Bibliographie

Cohen, G., et Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics*, 18, 13-23.

Draper, D. (1995). Assessment and propagation of model uncertainty (avec discussion). *Journal of the Royal Statistical Society, Series B*, 57, 45-97.

Smith, A.F.M., et Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46, 84-88.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

National Center for Health Statistics (1992). Third national health and nutrition examination survey. *Vital and Health Statistics, Series 2*, 113.

Nandram, B., Lin, N., Choi, J.W., et Cox, L. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.

Nandram, B., Han, G. et Choi, J.W. (2002). Un modèle bayésien hiérarchique de non-réponse non-ignorable pour les données multinomiales des petites régions. *Techniques d'enquête*, 28, 157-170.

Nandram, B., Cox, L.H. et Choi, J.W. (2005). Analyse bayésienne des données catégoriques manquantes non ignorables : une application à la densité minérale osseuse et au revenu familial. *Techniques d'enquête*, 31, 233-247.

Nandram, B., et Choi, J.W. (2005). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, 105, 120-135.

Nandram, B., et Choi, J.W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Techniques d'enquête*, 34, 41-54.

Nandram, B., et Choi, J.W. (2008). Une répartition bayésienne des électeurs indécis. *Techniques d'enquête*, 31, 79-92.

Nandram, B., et Choi, J.W. (2005). Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines : une application aux données de la NHANES. *Techniques d'enquête*, 31, 79-92.

Nandram, B., et Choi, J.W. (2004). A nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse. *Journal of Nonparametric Statistics*, 16, 821-839.

Nandram, B., et Choi, J.W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., et Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B. (2009). Bayesian inference of the cell probabilities of a two-way categorical table under non-ignorability. *Communications in Statistics - Theory and Methods*, 38, 3015-3030.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, deuxième édition. New York : John Wiley & Sons, Inc.

Kass, R.E., et Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Greenland, S. (2009). Relaxation penalties and priors for plausible modelling of nonidentified bias sources. *Statistical Sciences*, 24, 195-210.

Forster, J.J., et Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society, Series B*, 60, 57-70.

Geelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Greenland, S. (2009). Relaxation penalties and priors for plausible modelling of nonidentified bias sources. *Statistical Sciences*, 24, 195-210.

Kass, R.E., et Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, deuxième édition. New York : John Wiley & Sons, Inc.

Nandram, B. (2009). Bayesian inference of the cell probabilities of a two-way categorical table under non-ignorability. *Communications in Statistics - Theory and Methods*, 38, 3015-3030.

Nandram, B., et Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., et Choi, J.W. (2004). A nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse. *Journal of Nonparametric Statistics*, 16, 821-839.

Nandram, B., et Choi, J.W. (2005). Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines : une application aux données de la NHANES. *Techniques d'enquête*, 31, 79-92.

Nandram, B., et Choi, J.W. (2008). Une répartition bayésienne des électeurs indécis. *Techniques d'enquête*, 34, 41-54.

Nandram, B., Han, G. et Choi, J.W. (2002). Un modèle bayésien hiérarchique de non-réponse non-ignorable pour les données multinomiales des petites régions. *Techniques d'enquête*, 28, 157-170.

Nandram, B., Lin, N., Choi, J.W., et Cox, L. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.

National Center for Health Statistics (1992). Third national health and nutrition examination survey. *Vital and Health Statistics, Series 2*, 113.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Smith, A.F.M., et Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46, 84-88.

Statistique Canada, N° 12-001-X au catalogue

Notons que $0 \leq \pi_{\ell}^s \leq 1$, $\sum_{s=1}^4 \pi_{\ell}^s = 1$ et $0 \leq \phi_{\ell}^{jk} \leq 1$. Nous simplifions le calcul pour ℓ dans (A.3) en deux étapes. Premièrement, dans (A.3), nous effectuons la transformation

$$\phi_{ijk}^* = T_{ijk}^* \sum_c \sum_{k=1}^f \phi_{ijk}^* = T_{ij}^*.$$

[illegible]

où $b_i = \min \{ |I|/\Phi^{ijk} : f = 1, \dots, r; k = 1, \dots, c \}$. Deuxièmement, en posant que $W_i = \{\beta T_i / (1 - \pi_i)\}$ et en absorbant le facteur $\beta_{\text{csp}} / (\Gamma(\text{rcp}))$ dans I_i avec certaines opérations algébriques supplémentaires, nous obtenons

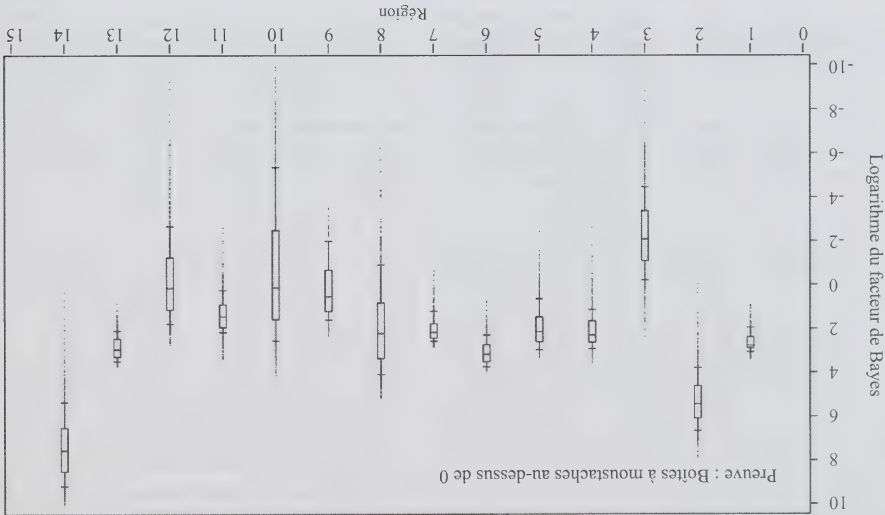
[illegible]

L'objectif du présent article était d'élaborer une méthodologie pour analyser les données provenant de tableaux incomplets de contingence à deux entrées, chaque tableau correspondant à un domaine ou région. Nous avons pour cela étendu la méthodologie bayésienne de Nandram et Choi (2002a, b) pour données binaires à des tableaux de contingences $r \times c$ pour petits domaines. Nous avons construit un nouveau modèle de non-réponse non ignorable bayésien (c'est-à-dire le modèle d'extension) qui est centré sur le modèle de non-réponse ignorable. Nous avons utilisé des méthodes Monte Carlo par chaîne de Markov (spécifiquement l'échantillonneur Metropolis-Hastings « gridy ») pour ajuster le modèle. Nous avons comparé notre modèle à un modèle de non-réponse ignorable et un modèle de non-réponse non ignorable. Enfin, nous avons illustré notre méthode en estimant les probabilités de cellule pour le tableau de contingence 3×3 de la densité minérale osseuse et du revenu sur 13 régions infranationales.

Nous avons montré qu'il existe des différences entre les trois modèles. En utilisant les données sur la densité minérale osseuse et le revenu familial, nous avons montré que notre modèle d'extension est un compromis entre le modèle de non-réponse ignorable et le modèle de non-réponse non ignorable. À l'aide du facteur de Bayes, nous avons montré qu'il existe des différences entre les tests de l'association de la densité minérale osseuse et du revenu familial quand les fréquences de cellule sont estimées au

5. Conclusion

Figure 1 Boîtes à moustaches des logarithmes du facteur de Bayes
L'association entre la DMO et le RF



Le fait que la preuve de l'association est nettement plus forte sous notre modèle que sous la procédure de ratiissage. En fait, grâce à l'emprunt d'information, nous avons pu obtenir l'information nécessaire. Deuxièmement, il serait également intéressant de relâcher l'hypothèse voulant que les marges du tableau de contingence soient fixes ; voir, par exemple, Nandram (2009) qui a procédé à cet examen pour une seule grande région. Troisièmement, le calage pourrait être encore amélioré (c'est-à-dire en intégrant de l'information sur les marges).

Remerciements

Nous exprimons notre reconnaissance à l'égard de deux examinateurs et d'un rédacteur associé pour l'aide apportée dans la présentation du document.

Annexe A

Densité a posteriori conjointe du modèle d'extension

Premièrement, en intégrant la fonction de densité a posteriori conjointe sur p , nous obtenons

non ignorable, nous avons un échantillon de 1 000 tableaux pour chaque région. Nous avons aussi un seul tableau combiné pour toutes les régions sous la procédure de ratisage et 1 000 tableaux pour toutes les régions combinées. Nous obtenons le facteur de Bayes pour chaque tableau sous un modèle multinomial-Dirichlet. Il convient de souligner que notre méthode s'appuie sur le modèle d'extension de sorte que les fréquences de cellule sont produites en empruntant de l'information à d'autres régions contrairement à la procédure de ratisage.

Puis, pour chaque tableau, nous prenons

$$n|\pi \sim \text{Multinomiale}(n, \pi) \text{ et } \pi \sim \text{Dirichlet}(1).$$

où $p_0(u) = n! \prod_{j=1}^J \pi_j^{n_j} (n + \nu)^{-n}$. Observons que $p_0(u)$ n'est pas une fonction de $\{n^k\}$. Donc, en tant que mesure d'association, c'est l'écart de $\prod_{j=1}^J n_j! \prod_{k=1}^K n^k$ par rapport à $\prod_{j=1}^J \prod_{k=1}^K n^k$ qui importe. Par contre, nous notons que, pour la statistique d'indépendance de Pearson classique, ce sont les écarts de n^k par rapport à $n_j n^k$ qui importent. Cependant, soulignons que ce test ne peut pas être appliqué, parce que bon nombre de fréquences de cellule prévues sont inférieures à 5 sous l'hypothèse de l'absence d'association et l'échantillonnage multinomial

Nous présentons nos résultats au tableau 4 et à la figure 1 qui correspond aux données du tableau 1 pour la classification croisée de la densité minérale osseuse et du revenu familial. Nous présentons les logarithmes des vraisemblances marginales (base e) et des facteurs de Bayes : ceux-ci doivent être interprétés en appliquant la règle empirique de Kass et Raftery (1995).

À la figure 1, nous voyons que les boîtes à moustaches se trouvent toutes au-dessus de zéro, sauf celle pour la troisième région qui ne donne aucune preuve d'association ; il n'existe peut-être pas non plus de preuve d'association pour la région 42 (10 dans la figure). Un résumé de ces résultats est présenté au tableau 4. Les facteurs de Bayes indiquent une association dans toutes les régions, sauf la région 12, et leur valeur réelle est nettement plus grande sous le modèle de non-réponse non ignorable. Les valeurs pour la région 6 et pour toutes les régions confondues sont élevées (336,3 c. 5,8 et 3 798,2 c. 0,183).

Tableau 4 Données de la NHANES pour 13 régions. Comparaison des vraisemblances marginales négatives et des facteurs de Bayes ou de l'association de la DMO et du RF d'après la procédure de ratisage et le modèle d'extension, selon la région

Région	$-\ln\{p_0(n)\}$	$-\ln\{p_u(n)\}$	FB	$-\ln\{p_u(n)\}$	FB	Extension
4	26,19	23,07	22,855	23,5014	14,78	1478,0169
6	45,73	43,98	5,766	40,5038	336,27	11,465
12	31,14	38,01	0,001	33,40,054	0,37	0,027
17	29,13	27,03	8,134	27,0026	10,27	0,191
25	25,44	26,02	0,558	23,8029	9,55	0,202
26	26,89	23,18	40,562	23,90,018	24,71	0,370
29	23,21	20,87	10,301	21,30,018	8,40	0,115
36	34,99	36,09	0,330	33,10,064	21,13	0,28
39	23,77	24,89	0,325	23,60,044	2,24	0,68
42	29,51	30,21	0,497	30,30,099	4,33	0,255
44	25,61	30,48	0,008	24,40,027	5,19	0,137
48	38,83	35,34	32,650	39,10,060	2,15	0,081
53	27,11	24,82	9,865	24,20,017	19,40	0,282
All	53,43	55,13	0,183	46,10,049	3,798,24	151,82

Nota : Dans la colonne des régions, « Toutes » désigne toutes les régions confondues ; la notation u_b signifie que la moyenne est u et que l'erreur-type est b sur les 1 000 itérations. Le terme $\ln\{p_0(n)\}$ est le même pour les deux procédures.

Tableau 3
Moyennes a posteriori des probabilités de cellule et des intervalles de crédibilité (IC) à 95 % pour trois régions (grande, moyenne et petite) selon les trois modèles

Cellule	Modèle 1			Modèle 2			Modèle 3		
	MP	ETP	IC à 95 %	MP	ETP	IC à 95 %	MP	ETP	IC à 95 %
a. Grande	0.239	0.044	(0.157; 0.326)	0.196	0.046	(0.117; 0.295)	0.259	0.038	(0.189; 0.335)
(1.1)	0.140	0.035	(0.078; 0.213)	0.127	0.035	(0.068; 0.200)	0.132	0.029	(0.082; 0.197)
(1.2)	0.240	0.044	(0.159; 0.332)	0.198	0.047	(0.118; 0.301)	0.248	0.037	(0.175; 0.322)
(1.3)	0.092	0.032	(0.039; 0.162)	0.098	0.040	(0.037; 0.188)	0.077	0.022	(0.024; 0.126)
(2.1)	0.074	0.028	(0.029; 0.136)	0.077	0.030	(0.030; 0.144)	0.056	0.020	(0.024; 0.099)
(2.2)	0.133	0.036	(0.070; 0.210)	0.121	0.042	(0.056; 0.219)	0.110	0.028	(0.058; 0.168)
(2.3)	0.036	0.020	(0.008; 0.083)	0.069	0.039	(0.013; 0.153)	0.047	0.018	(0.018; 0.086)
(3.1)	0.023	0.015	(0.003; 0.061)	0.043	0.025	(0.007; 0.100)	0.032	0.014	(0.009; 0.063)
(3.2)	0.025	0.015	(0.003; 0.061)	0.043	0.025	(0.007; 0.100)	0.032	0.014	(0.009; 0.063)
(3.3)	0.025	0.017	(0.003; 0.066)	0.071	0.040	(0.010; 0.154)	0.042	0.016	(0.016; 0.079)
b. Moyenne	0.233	0.034	(0.169; 0.302)	0.213	0.043	(0.141; 0.305)	0.254	0.032	(0.194; 0.318)
(1.1)	0.143	0.028	(0.093; 0.200)	0.127	0.032	(0.072; 0.196)	0.146	0.024	(0.102; 0.197)
(1.2)	0.190	0.031	(0.132; 0.254)	0.140	0.034	(0.084; 0.218)	0.208	0.027	(0.156; 0.259)
(1.3)	0.174	0.031	(0.118; 0.237)	0.160	0.042	(0.092; 0.249)	0.154	0.027	(0.106; 0.211)
(2.1)	0.043	0.018	(0.015; 0.083)	0.060	0.028	(0.017; 0.124)	0.032	0.012	(0.012; 0.059)
(2.2)	0.049	0.020	(0.017; 0.095)	0.065	0.031	(0.018; 0.136)	0.042	0.014	(0.020; 0.072)
(2.3)	0.112	0.025	(0.068; 0.167)	0.120	0.041	(0.059; 0.209)	0.092	0.020	(0.056; 0.134)
(3.1)	0.047	0.018	(0.018; 0.088)	0.059	0.026	(0.019; 0.118)	0.040	0.014	(0.018; 0.071)
(3.2)	0.087	0.042	(0.022; 0.184)	0.107	0.065	(0.013; 0.144)	0.103	0.043	(0.012; 0.086)
(3.3)	0.087	0.042	(0.022; 0.184)	0.107	0.065	(0.013; 0.144)	0.103	0.043	(0.012; 0.086)
c. Petite	0.196	0.052	(0.103; 0.305)	0.164	0.055	(0.077; 0.288)	0.253	0.043	(0.175; 0.334)
(1.1)	0.081	0.034	(0.028; 0.158)	0.081	0.032	(0.030; 0.155)	0.091	0.028	(0.043; 0.152)
(1.2)	0.213	0.052	(0.118; 0.323)	0.175	0.055	(0.087; 0.300)	0.220	0.043	(0.137; 0.306)
(1.3)	0.093	0.041	(0.028; 0.186)	0.111	0.055	(0.029; 0.234)	0.073	0.028	(0.030; 0.139)
(2.1)	0.056	0.029	(0.012; 0.126)	0.066	0.031	(0.018; 0.136)	0.045	0.020	(0.014; 0.094)
(2.2)	0.115	0.045	(0.042; 0.215)	0.118	0.053	(0.038; 0.240)	0.092	0.030	(0.041; 0.158)
(2.3)	0.048	0.028	(0.036; 0.222)	0.113	0.056	(0.031; 0.239)	0.081	0.030	(0.033; 0.148)
(3.1)	0.044	0.028	(0.006; 0.113)	0.065	0.035	(0.013; 0.144)	0.043	0.020	(0.012; 0.086)
(3.2)	0.042	0.028	(0.022; 0.184)	0.107	0.065	(0.013; 0.144)	0.103	0.043	(0.012; 0.086)
(3.3)	0.042	0.028	(0.022; 0.184)	0.107	0.065	(0.013; 0.144)	0.103	0.043	(0.012; 0.086)

Nota : Voir l'annexe B pour une description des modèles 1 et 2.

4.3 Facteur de Bayes pour la preuve d'association

Nous avons également considéré l'association entre la densité minérale osseuse et le revenu familial. Bien que l'existence d'une telle association paraisse peu probable, il est intéressant d'examiner cette question ; voir Nandram, Cox et Choi (2005) pour une discussion de ce problème. Nous nous servons du facteur de Bayes (Kass et Raftery 1995) pour mesurer la force de la preuve d'une association comparative d'absence d'association dans le tableau de contingence $r \times c$. Nous le faisons pour chacune des 13 régions et pour toutes les régions confondues.

Nous utilisons deux procédures, l'une sans modélisation étendue et l'autre s'appuyant sur notre modèle (étendu) de non-réponse non ignorable. La méthode simple consiste à produire les fréquences de cellule selon une technique de ratisage ordinaire, et nous supposons qu'il n'y a pas d'erreur à le faire. Il s'agit d'une procédure dicée par le bon sens que les praticiens des enquêtes utilisent régulièrement. Au moyen de la deuxième procédure fondée sur notre modèle de non-réponse non ignorable, nous avons obtenu 1 000 tableaux combinés pour chaque région, comme il est

décrit à la section 3 sur les calculs. Pour chaque région, nous avons obtenu les fréquences de cellule pour les quatre tableaux, et nous les avons totalisées pour obtenir un seul tableau de toutes les fréquences. Voici la description de la procédure de ratisage pour obtenir les fréquences de cellule. Soit $n_{jk}^{(1)}$ les fréquences de cellule pour les quatre tableaux combinés. Soit $n_{jk}^{(1)}$ les fréquences de cellule pour le tableau des totaux de données complètes, $n_{jk}^{(2)}$ celles pour le tableau des totaux de ligne, $n_{jk}^{(3)}$ celles pour le tableau des totaux de colonne et $n_{jk}^{(4)}$ celles pour le tableau des totaux. Les fréquences de cellule pour les quatre tableaux sont estimées par

$$n_{jk}^{jk} = n_{jk}^{(1)} + \left(\frac{n_{jk}^{(2)}}{n_{(1)}^{(2)}} \right) n_{j,c+1}^{(2)} + \left(\frac{n_{jk}^{(3)}}{n_{(1)}^{(3)}} \right) n_{r+1,k}^{(3)} + \left(\frac{n_{jk}^{(4)}}{n_{(1)}^{(4)}} \right) n_{r+1,c+1}^{(4)}$$

$j = 1, \dots, r, k = 1, \dots, c.$

Dans chaque cas, nous désignons la somme des fréquences de cellule pour chaque région par n_{jk}^{jk} . Pour la procédure de ratisage, nous n'avons qu'un seul tableau pour chaque région, tandis que pour le modèle de non-réponse

Techniques d'enquête, juin 2012

respectivement, ce qui représente une bonne concordance. Les estimations de τ_1 et τ_2 devraient présenter les écarts les plus importants, mais elles sont également raisonnables pour τ_1 , l'intervalle de crédibilité à 95 % est (28,282 ; 64,204) avec l'approximation et (27,962 ; 64,425) avec l'algorithme SIR]. Dans les deux cas, les ETN sont faibles, ce qui signifie que les calculs sont répétables.

Au tableau 3, nous comparons notre modèle d'extension (modèle 3) à deux autres. Le modèle 1 (un modèle de non-réponse ignorable) et le modèle 2 (un modèle de non-réponse non ignorable) (pas de centrage) sont décrits à qu'elles sont toutes inférieures à 0,005.

Tableau 1
Fréquences des tableaux 3 × 3 de la DMO et du RF correspondant aux 13 régions infranationales dans la NHANES III

État	Tableau complet			Total de colonne			Total de ligne			Total
6	21	14	9	8	7	32	51	92	106	257
12	33	18	21	4	4	15	22	18	33	127
17	25	15	13	6	5	3	12	6	17	65
25	9	7	12	6	5	9	4	2	17	28
26	18	11	18	6	5	9	4	3	17	4
29	9	4	10	3	2	4	0	2	4	0
36	42	17	27	13	18	9	6	7	42	9
39	8	6	14	2	5	4	3	2	9	0
42	14	8	11	12	8	4	3	1	12	0
44	12	9	6	8	5	0	1	1	7	0
48	159	44	22	51	11	13	9	8	16	2
53	14	10	15	10	10	14	3	1	8	1

Nota : Dans le tableau 3 × 3 complet, le premier (deuxième, troisième) ensemble de trois nombres correspond à la première (deuxième, troisième) ligne ; le total de colonne (ligne) renvoie au tableau 3 × 3 ne contenant que les totaux de colonne (ligne) ; le total renvoie au tableau 3 × 3 contenant les totaux uniquement.

Tableau 2
Données de la NHANES pour 13 régions : Comparaison de la densité de probabilité a posteriori approximative et de la densité de probabilité a posteriori correcte en utilisant les moyennes a posteriori (MP), les écarts-types a posteriori (ETP), les erreurs-types numériques (ETN) et les intervalles de crédibilité à 95 % des hyperparamètres

Approximation			Corrigée		
MP	ETP	ETN	MP	ETP	ETN
0,528	0,031	0,001	0,525	0,031	0,008
0,131	0,021	0,001	0,133	0,021	0,002
0,328	0,028	0,001	0,328	0,028	0,005
0,013	0,006	0,000	0,014	0,006	0,000
21,638	9,559	0,255	20,078	8,632	0,303
0,280	0,023	0,001	0,277	0,023	0,004
0,133	0,016	0,000	0,134	0,019	0,002
0,200	0,019	0,000	0,199	0,017	0,003
0,105	0,015	0,000	0,107	0,015	0,002
0,065	0,011	0,000	0,065	0,011	0,001
0,072	0,012	0,000	0,073	0,012	0,001
0,061	0,011	0,000	0,061	0,011	0,001
0,037	0,008	0,000	0,036	0,008	0,001
0,048	0,009	0,000	0,048	0,009	0,001
45,960	10,094	0,153	45,177	10,562	0,679
1,472	0,218	0,004	1,449	0,208	0,022

Nota : Les hyperparamètres sont $\mu_1, \mu_2, \tau_1, \tau_2$ et β .

4. Un exemple

En guise d'illustration, nous choisissons un exemple dans le domaine de la statistique de la santé. À la section 4.1, nous décrivons brièvement les données provenant de la troisième édition de la National Health and Nutrition Examination Survey (NHANES III) que nous utilisons. En particulier, nous étudions la relation entre la densité minérale osseuse et le revenu familial ; voir Nandram, Cox et Choi (2005) pour une discussion de ce problème. À la section 4.2, après une brève discussion de nos calculs, nous présentons l'inférence a posteriori sur les probabilités de cellule. À la section 4.3, en nous servant du facteur de Bayes, nous discutons de la relation entre la densité minérale osseuse et le revenu familial.

4.1 Données de la NHANES III

Le plan de sondage est un plan probabiliste stratifié à plusieurs degrés qui est représentatif de l'ensemble de la population civile ne vivant pas en établissement, âgée de deux mois et plus, des États-Unis. Des renseignements plus détaillés sur le plan de sondage de la NHANES III sont disponibles ailleurs (National Center for Health Statistics 1992, 1994). La collecte des données de la NHANES III comporte deux volets : le premier comprend la sélection de l'échantillon et l'interview des membres des ménages échantillonnés en vue de recueillir des renseignements personnels et le second comprend l'examen physique des personnes interviewées dans un centre d'examen mobile (CEM). L'évaluation de la santé s'appuie sur un examen physique, des tests et des mesures faites par des techniciens, ainsi que des prélèvements pour l'analyse. L'échantillon a été sélectionné auprès des ménages de 81 unités primaires d'échantillonnage à travers les États-Unis contenantau d'octobre 1988 à septembre 1994. Les données finales retenues pour l'étude proviennent des 35 plus grandes unités primaires d'échantillonnage dont la population est égale ou supérieure à 500 000 habitants, et nous considérons 13 régions infranationales.

La non-réponse peut avoir lieu dans les volets interview et examen physique de l'enquête. La non-réponse à l'interview se produit lorsque les personnes échantillonnées ne participent pas à l'interview. Certaines personnes interviewées et incluses dans le sous-échantillon pour l'évaluation de la santé ont manqué l'examen physique à la maison ou au centre d'examen mobile et n'ont donc pas subi la totalité ou une partie des examens.

Les médecins pensent que les personnes obèses ou ayant un excès de poids ne se présentent généralement pas au CEM. Cohen et Duffy (2002) remarquent que les enquêtes sur la santé sont un bon exemple de situation où il paraît plausible qu'il existe un lien entre la propension à répondre

et l'état de santé. La NHANES III est en effet un bon exemple.

Les personnes échantillonnées pour participer à la NHANES III peuvent être classées en fonction d'un grand nombre d'attributs, et les chercheurs analysent ces tableaux de contingence afin de déterminer la qualité de l'ajustement des modèles ou l'indépendance. Ici, nous étudions la densité minérale osseuse (DMO) et le revenu familial (RF). Men-tionnons ici que, même si le RF est une variable discrète, nous avons classé la DMO en trois catégories (normale, ostéopénie et ostéoporose) et le RF en trois catégories (faible, moyen et élevé). Cependant, nous ne disposons que d'une classification partielle des individus, parce que certains sont classés en fonction d'un seul attribut, tandis que d'autres ne sont pas classés du tout. Parmi les ménages qui ont participé au volet de l'examen physique, environ 62 % ont fourni des données sur le RF et la DMO, 8 % ont fourni des données sur le revenu seulement et 1 % n'ont fourni de données ni sur le revenu ni sur la DMO. Notre problème consiste à estimer les probabilités de cellule et de tester l'association entre la DMO et le RF pour chacune des 13 régions infranationales en utilisant notre modèle d'extension qui regroupe les données de manière adaptative. Dans le tableau 1, nous présentons les tableaux 3×3 de la DMO et du RF pour les 13 régions susmentionnées. Notons que les données pour les régions 6 et 48 sont suffisantes pour traiter ces régions individuellement. Par contre, les autres régions sont très petites. Les fréquences dans le tableau contenant des totaux de ligne sont généralement faibles, sauf pour la région 17, et les fréquences dans le tableau contenant juste le total sont faibles. Même pour le cellule sont généralement faibles, ce qui nous oblige à emprunter de l'information.

4.2 Inférence a posteriori des probabilités de cellule

Nous discutons de la performance de nos calculs pour le modèle d'extension, puis de l'inférence a posteriori au sujet des probabilités de cellule. Nous utilisons la moyenne a posteriori (MP), l'écart-type a posteriori (ETP) et l'intervalle de crédibilité à 95 % pour chaque paramètre d'intérêt. Nous présentons aussi les erreurs-types numériques (ETN) pour évaluer la répétabilité de nos calculs.

Au tableau 2, nous présentons des résumés des distributions a posteriori de μ_i , π_i , τ_i et β_i , avant et après l'application de l'algorithme SIR. Ces résumés sont fort semblables, ce qui indique que l'approximation $\pi_{(S, Y^{(1)})}$, $Y_i^{(1)}, u_i, v_i, w_i$ donnée par l'algorithme SIR n'est pas déraisonnable. Par exemple, les intervalles de crédibilité à 95 % obtenus pour β avant et après l'application de l'algorithme SIR sont (1,081 ; 1,940) et (1,086 ; 1,947),

ou

$$I_j = \iiint_{\beta_j, \tau_1, \tau_2} G^{\tau_2 \beta_j} \left(\frac{\beta_j}{1 - \pi_{i1}} \right) \left[G^{\beta_j} \left(\frac{\beta_j}{1 - \pi_{i1}} \right) \right]$$

$$\prod_{j,k} \left\{ \left(\frac{\beta_j}{W_i} \right)^{\frac{1}{4}} \sum_{s=2}^{J_{ijk}} \left[\frac{1}{1 - \pi_{i1}} \right]^{\frac{\beta_j}{4}} \left\{ 1 - \frac{\pi_{i1}}{\beta_j} \right\} \right\} \left[\right]_{Y_{ijk}}$$

$$W_i^{\tau_2 \beta_j - 1} e^{-W_i} \frac{\Gamma(\tau_2 \beta_j) G^{\tau_2 \beta_j} \left(\frac{\beta_j}{1 - \pi_{i1}} \right) \left(\frac{\beta_j}{1 - \pi_{i1}} \right) \left\{ D(Y_{i1}^{(1)}) + \beta_j \right\}}{\prod_{j,k} \Phi_{Y_{ijk}, \mu + \beta_j - 1}^*} \left\{ \prod_{s=1}^4 \left\{ D(Y_{i2}^{(2)}) + \mu + \tau_2 \right\} \right\} \Phi^* D \pi_{i1}$$

avec $\beta_j = \min \{1 / \Phi_{ijk}^*\}$, $j = 1, \dots, r$, $k = 1, \dots, c$ et

$$= \frac{a_0}{b_0} \cdot \frac{(a_0 + \tau_1)^2}{(b_0 + \tau_2)^2} \cdot \frac{(c_0 + \beta_j)^2}{c_0} = \frac{\prod_{j=1}^4 \frac{\Gamma(\tau_2 \beta_j)}{\Gamma(\beta_j)} D(Y_{i1}^{(1)}) + \beta_j}{n_i} \times \prod_{j=1}^4 \left(Y_{i1}^{(1)}, Y_{i2}^{(2)}, Y_{i3}^{(3)}, Y_{i4}^{(4)} \right) \frac{D(\mu_1 \tau_1)}{D(Y_{i1}^{(1)} + \mu_1 \tau_1)} \frac{D(\mu_2 \tau_2)}{D(Y_{i2}^{(2)} + \mu_2 \tau_2)} \quad (14)$$

Pour évaluer I_i pour chaque $i = 1, \dots, A$, nous procédons comme il suit sachant $(\Omega, Y_{i1}^{(1)})$:

1. Tirer des échantillons indépendants des vecteurs π_i et Φ_i^* des distributions Dirichlet $(j_i^{(2)} + \mu_2 \tau_2)$ et Φ_i^* des distributions Dirichlet $(j_i^{(1)} + \beta_j)$, tirer un échantillon des valeurs de W_i de la distribution gamma tronquée sur l'intervalle $(0, \{\beta_j/1 - \pi_{i1}\})$ avec le paramètre $\tau_2 \beta_j$.

2. Pour chaque π_i, Φ_i^* et W_i sélectionné à l'étape (1), calculer R_1, R_2 , ou

$$R_1 = G^{\tau_2 \beta_j} \left(\frac{\beta_j}{1 - \pi_{i1}} \right) \left[G^{\beta_j} \left(\frac{\beta_j}{1 - \pi_{i1}} \right) \right] \quad (15)$$

$$R_2 = \prod_{j,k} \left(\frac{\beta_j}{W_i} \right)^{\frac{1}{4}} \left\{ \prod_{s=2}^{J_{ijk}} \left[\frac{1}{1 - \pi_{i1}} \right]^{\frac{\beta_j}{4}} \left\{ 1 - \frac{\pi_{i1}}{\beta_j} \right\} \right\} \left[\right]_{Y_{ijk}} \quad (16)$$

3. Répéter les étapes (1) et (2) 1 000 fois. Puis, calculer la moyenne de R_1, R_2 sur ces 1 000 valeurs.

La suite des calculs comporte deux parties. Premièrement, nous utilisons l'échantillonneur de Metropolis-Hastings « griddy » (technique d'approximation par des grilles) pour tirer des échantillons de $\pi_a(\Omega, Y^{(1)} | Y^{(1)}, n, v, w)$. Nous échantillons μ_1, μ_2, τ_1 et τ_2 à partir de leur densité des probabilité a posteriori conditionnelle en utilisant des grilles; cela comporte la transformation de τ_1 et τ_2 pour les faire varier dans l'intervalle unitaire (0, 1). Pour chaque distribution, nous utilisons 100 grilles; voir Nandram, Cox et Choi (2005) pour une procédure similaire. Ici, le tirage de $Y^{(1)}$ est effectué par échantillonnage de la fonction de masse de probabilité conditionnelle par composante. Des échantillons sont tirés de la densité a posteriori conditionnelle de β en utilisant un pas de Metropolis de manière similaire à Nandram et Choi (2002a, b). Nous avons exécuté cet algorithme 11 000 fois en permettant un « rodage » de 1 000 itérations. Nous avons constaté que les autocorrélations entre les itérations étaient faibles, ce qui indiquait que l'échantillonneur produisait un bon mélange. Nous avons également utilisé la méthode des moyennes de lot pour étendre l'évaluation des calculs. Nous avons utilisé des lots de 25 pour calculer les erreurs-types numériques. Deuxièmement, nous servons de l'algorithme SIR pour sous-échantillonner l'échantillon de 10 000 itérations tirées de $\pi_a(\Omega, Y^{(1)} | Y^{(1)}, n, v, w)$. Pour chacune des 10 000 itérations, nous calculons les poids

$$w^m = \frac{h(\Omega_{(m)}, Y_{(m)}^{(1)} | Y^{(1)}, n, v, w)}{\pi_a(\Omega_{(m)}, Y_{(m)}^{(1)} | Y^{(1)}, n, v, w)}, \quad m = 1, \dots, M = 10\,000 \quad (17)$$

Enfin, nous pouvons maintenant faire une inférence exacte (dans les limites des méthodes Monte Carlo par chaîne de Markov) au sujet de D_i a posteriori. Soit $Y_{i,jk} = \sum_{s=1}^4 Y_{ijk}^{(s)}$. Alors, $D_i | Y_i^*, \mu_1, \tau_1 \sim \text{Dirichlet}(Y_i^* + \mu_1 \tau_1)$, $i = 1, \dots, A$. Donc, pour chaque valeur de Y_i^*, μ_1 et τ_1 , que nous obtenons au moyen de l'algorithme SIR, nous tirons une valeur de $D_i, i = 1, \dots, A$. D'où nous obtenons une densité blackwellisée pour chaque D_i , et l'inférence se poursuit de la manière habituelle.

$$\mathcal{G}_{\mathcal{Z}_i}(\phi | \beta, \pi_i)$$

$$= \left\{ \prod_{j=1}^J \prod_{k=1}^K \left[\frac{\beta^{\frac{\phi_{j,k}}{\beta}-1} e^{-\frac{1-\pi_{ij}}{\beta\phi_{j,k}}}}{\beta^{\frac{\phi_{j,k}}{\beta}} \Gamma(\beta) G[\beta(1-\pi_{ij})^{-1}]} \right] \right\}, \quad (9)$$

pour $0 < \phi_{j,k}^{\frac{\phi_{j,k}}{\beta}} < 1$. La loi priori conjointe de π_i et ϕ_i est simplement le produit de $\mathcal{G}_{\mathcal{Z}_i}(\phi_i | \beta, \pi_i)$ et de $\mathcal{G}_{\mathcal{Z}_i}(\pi_i | \mu_2, \tau_2)$. Donc, la loi a priori conjointe de $\phi = (\phi_1, \dots, \phi_A)'$ et de π est

$$\mathcal{G}^*(\pi, \phi | \mu_2, \tau_2, \beta) = \prod_{i=1}^I \{\mathcal{G}_{\mathcal{Z}_i}(\phi_i | \beta, \pi_i) \cdot \mathcal{G}_{\mathcal{Z}_i}(\pi_i | \mu_2, \tau_2)\}.$$

Autrement dit

$$\mathcal{G}^*(\pi, \phi | \mu_2, \tau_2, \beta)$$

$$= \left\{ \prod_{i=1}^I \prod_{j=1}^J \prod_{k=1}^K \frac{\prod_{c=1}^C \pi_{i,j,k}^{\mu_{2,j,k}-1}}{\mathcal{D}(\mu_2, \tau_2)} \right\}$$

$$\left\{ \frac{\beta^{\frac{\phi_{j,k}}{\beta}-1} e^{-\frac{1-\pi_{ij}}{\beta\phi_{j,k}}}}{\beta^{\frac{\phi_{j,k}}{\beta}} \Gamma(\beta) G[\beta(1-\pi_{ij})^{-1}]} \right\}. \quad (10)$$

Pour achever la description du modèle, nous spécifions les hypothèses concernant les hyperparamètres. Comme il n'existe pas de lois a priori conjuguées, nous utilisons des lois a priori de rétrécissement pour τ_1, τ_2 et β , parce qu'elles sont propres et non informatives. Les lois a priori de la forme $p(\tau_1) \propto 1/\tau_1$, en particulier les lois a priori gamma diffusées, sont déconseillées; voir, par exemple, Gelman (2006). Des demi-densités de la loi de Cauchy et des densités de probabilité de la loi gamma sont d'autres options (pour lesquelles il faudrait spécifier les hyperparamètres). Donc, nous prenons

1. τ_1, τ_2 et β ayant des lois a priori de rétrécissement indépendantes de la forme

$$f(x) = \frac{a_0}{a_0 + x} z^{\frac{a_0 + x}{z}}, \text{ pour } x \geq 0,$$

où a_0 est spécifié; il est courant en pratique de prendre $a_0 = 1$.

2. Nous supposons aussi que $\mu_1 \sim \text{Dirichlet}(1, 1, \dots, 1)$ et $\mu_2 \sim \text{Dirichlet}(1, 1, 1)$.

Soit $\Omega = (\beta, \mu_1, \tau_1, \mu_2, \tau_2)$. La densité de probabilité pour Ω est alors

$$p(\Omega) = \frac{a_0}{b_0} \cdot \frac{c_0 + \tau_1}{b_0} \cdot \frac{c_0 + \tau_2}{c_0} \cdot \frac{c_0}{c_0} (c_0 - 1)! z^{\frac{c_0}{c_0}}$$

pour τ_1, τ_2 et $\beta \geq 0$, $\sum_{j,k} \mu_{1,j,k} = 1$ et $\sum_{s=1}^S \mu_{2,s} = 1$.

3. Calculs

Pour faire des inférences au sujet des P^{ijk} , nous tirerons des échantillons de $h(\Omega, p, \pi, \phi, y^{(1)} | y_1, n, v, w)$ selon des méthodes Monte Carlo par chaîne de Markov. Cette procédure est décrite à la section 3.

$$a_{ijk}(\pi_{i1}, \phi_{ijk}) = \left(\frac{\phi_{ijk}}{1 - \pi_{i1}} \right) \left[1 + \frac{\pi_{i1}}{1} \{1 - \pi_{i1} - \phi_{ijk}\} \right]. \quad (12)$$

où, en substituant $(1 - \pi_{i1})^{-1} \phi_{ijk}$ pour ψ_{ijk} ,

$$\times \frac{a_0}{b_0} \cdot \frac{c_0 + \tau_1}{b_0} \cdot \frac{c_0 + \tau_2}{c_0} z^{\frac{c_0}{c_0}} \quad (11)$$

$$\left\{ \frac{\beta^{\frac{\phi_{j,k}}{\beta}-1} e^{-\frac{1-\pi_{ij}}{\beta\phi_{j,k}}}}{\beta^{\frac{\phi_{j,k}}{\beta}} \Gamma(\beta) G[\beta(1-\pi_{ij})^{-1}]} \right\}$$

$$\times \left\{ \frac{\mathcal{D}(\mu_1, \tau_1)}{\prod_{h=1}^H \prod_{k=1}^K \pi_{1,h,k}^{\mu_{1,h,k}-1}} \right\}$$

$$\left\{ \prod_{j=1}^J \prod_{k=1}^K \pi_{i,j,k}^{\pi_{i,j,k}-1} \times \prod_{j=1}^J \prod_{k=1}^K \{P^{ijk} \phi_{ijk}^{\phi_{ijk}}\} \right\} [a_{ijk}(\pi_{i1}, \phi_{ijk})]^{y_{i1,i}}$$

$$= \left(\prod_{j=1}^J \prod_{k=1}^K \left(y_{i1,i}^{y_{i1,i}} y_{i2,j}^{y_{i2,j}} y_{i3,k}^{y_{i3,k}} y_{i4}^{y_{i4}} \right) (1 - \pi_{i1})^{-y_{i1,i}} \right)$$

$$f(y | \pi, \phi, d, \mathcal{G}_1(d | \mu_1, \tau_1, \beta) \mathcal{G}^*(\pi, \phi | \mu_2, \tau_2, \beta) p(\Omega) h(\Omega, p, \pi, \phi, y^{(1)} | y_1, n, v, w) \propto$$

En vertu du théorème de Bayes, la densité de probabilité a posteriori conjointe est

Nous utilisons l'algorithme SIR pour sous-échantillonner un échantillon aléatoire tiré d'une densité de probabilité a posteriori approximative. L'exécution de cette tâche se fait en trois étapes. Nous agrégeons sur les P^{ij} , π_i et ϕ_i approximons la densité de probabilité agrégee par une densité plus simple et effectuons l'échantillonnage à partir de cette densité, puis nous sous-échantillonons ces échantillons pour obtenir des échantillons tirés de la densité de probabilité originale. Nous montrons à la présente section comment exécuter ces trois étapes.

Pour obtenir l'approximation et pour simplifier les calculs, à l'annexe A, nous procédons à l'agrégation sur les P^{ij} , π_i et ϕ_i pour obtenir

$$h(\Omega, y^{(1)} | y_1, n, v, w) = \pi_n(\Omega y^{(1)} | y_1, n, v, w) \cdot \prod_{j=1}^J \prod_{k=1}^K I_{j,k}$$

ou

$$y_{jrc}^{ts} = \text{vec}(\{y_{jrk}^{ts} \mid j = 1, \dots, r; k = 1, \dots, c\}),$$

$$y = (y_{11}^{t1}, y_{12}^{t1}, y_{13}^{t1}, y_{14}^{t1}, y_{21}^{t1}, y_{22}^{t1}, y_{23}^{t1}, y_{24}^{t1}, y_{31}^{t1}, y_{32}^{t1}, y_{33}^{t1}, y_{34}^{t1}, y_{41}^{t1}, y_{42}^{t1}, y_{43}^{t1}, y_{44}^{t1})',$$

$$\pi_{14 \times 1} = (\pi_{11}^{t1}, \dots, \pi_{14}^{t1}, \pi_{21}^{t1}, \dots, \pi_{24}^{t1}, \pi_{31}^{t1}, \dots, \pi_{34}^{t1})',$$

$$\psi_{4rc \times 1} = (\psi_{111}^{t1}, \dots, \psi_{114}^{t1}, \psi_{121}^{t1}, \dots, \psi_{124}^{t1}, \psi_{211}^{t1}, \dots, \psi_{214}^{t1}, \psi_{221}^{t1}, \dots, \psi_{224}^{t1}, \psi_{311}^{t1}, \dots, \psi_{314}^{t1}, \psi_{321}^{t1}, \dots, \psi_{324}^{t1}, \psi_{411}^{t1}, \dots, \psi_{414}^{t1})',$$

et

$$p_{4rc \times 1} = (p_{11}, p_{21}, \dots, p_{44})',$$

En obtenant des facteurs qui sont des puissances de π_{ts}^s , la fonction de vraisemblance peut également être exprimée sous la forme

$$f(y|\pi, p, d) = \prod_{j=1}^r \prod_{k=1}^c \left(y_{jrk}^{t1} \cdot y_{jrk}^{t2} \cdot y_{jrk}^{t3} \cdot y_{jrk}^{t4} \right)^{n_{jrk}^{t1}}$$

$$\left\{ \prod_{j=1}^r \prod_{k=1}^c \pi_{jrk}^{n_{jrk}^{t1}} \times \prod_{j=1}^r \prod_{k=1}^c (d_{jrk}^{t1} \psi_{jrk}^{t1})^{n_{jrk}^{t1}} \right\} \quad (6)$$

où $0 \leq \pi_{ts}^s \leq 1$, $\sum_s \pi_{ts}^s = 1$ et $0 \leq \psi_{jrk}^{ts} \leq (1 - \pi_{11}^{t1})^{-1}$. Ici, nous notons que y_{jrk}^{ts} et y_{jrk}^{t1} sont des variables observées, mais que les y_{jrk}^{ts} sont des variables latentes.

2.2 Construction des lois a priori

Les hypothèses qui suivent décrivent les lois a priori pour le modèle de non-réponse non ignorable :

1. Pour le vecteur de probabilités de cellule p_i , nous supposons que

$$p_i | \pi_i, \tau_i \sim \text{Dirichlet}(\pi_i, \tau_i),$$

où $\mu_1 = (\mu_{11}^{t1}, \mu_{12}^{t1}, \dots, \mu_{14}^{t1}, \mu_{21}^{t1}, \dots, \mu_{24}^{t1}, \mu_{31}^{t1}, \dots, \mu_{34}^{t1}, \mu_{41}^{t1}, \dots, \mu_{44}^{t1})$; $\mu_1^{trc} \geq 0$ et $\sum_{j=1}^r \sum_{k=1}^c \mu_{jrk}^{trc} = 1$. Le paramètre τ_1 nous informe de la similarité entre les p_i : plus la valeur de τ_1 est grande, plus les p_i se ressemblent. Il en est ainsi parce qu'une grande valeur de τ_1 signifie que les variances des p_i sont faibles, et comme elles ont la même moyenne, cela signifie qu'elles sont plus semblables quand τ_1 est grand.

Donc, pour p_i , la densité de probabilité est

$$g_i^1(p_i | \pi_i, \tau_i) = \prod_{j=1}^r \prod_{k=1}^c g_{ijk}^1(p_i | \pi_i, \tau_i)$$

$$= \prod_{j=1}^r \prod_{k=1}^c \frac{D(\mu_{jrk}^{t1})}{\prod_{j=1}^r \prod_{k=1}^c d_{jrk}^{t1}}$$

où, pour un k-uplet c et un scalaire t

$$D(ct) = \frac{\Gamma(t)}{\prod_{j=1}^k \Gamma(c_j t)}$$

(7)

2. Indépendamment des p_i , les $\pi_i = (\pi_{11}^{t1}, \pi_{12}^{t1}, \pi_{13}^{t1}, \pi_{14}^{t1})'$ suivent la spécification

$$\pi_i \sim \text{Dirichlet}(\mu_i, \tau_i),$$

avec $\pi_{ts}^s \geq 0$ et $\sum_s \pi_{ts}^s = 1$, où $\mu_i = (\mu_{i1}^{t1}, \mu_{i2}^{t1}, \mu_{i3}^{t1}, \mu_{i4}^{t1})'$, $\mu_{ts}^s \geq 0$, $\sum_{s=1}^4 \mu_{ts}^s = 1$ et τ_i est une mesure de la similarité entre les π_i . Donc, pour π_i , la densité de probabilité est

$$g_i^{2t}(\pi_i | \mu_i, \tau_i) = \frac{D(\mu_i, \tau_i)}{\prod_{j=1}^4 \mu_{j2}^{t2-1}} \quad (8)$$

3. Pour chaque i , soit $\psi_i = (\psi_{i1}^{t1}, \dots, \psi_{i4}^{t1}, \psi_{i1}^{t2}, \dots, \psi_{i4}^{t2}, \dots, \psi_{i1}^{trc}, \dots, \psi_{i4}^{trc})'$ de sorte que $\psi = (\psi_1^{t1}, \dots, \psi_4^{t1}, \psi_1^{t2}, \dots, \psi_4^{t2}, \dots, \psi_1^{trc}, \dots, \psi_4^{trc})'$. Nous supposons que, pour chaque i , les ψ_{ijk}^{ts} sont indépendants et identiquement distribués suivant une distribution

gamma ordinaire est tronquée comme il suit

$$\psi_{ijk} | \beta_i, \pi_i \sim \text{Gamma}(\beta_i, \psi_{ijk})$$

de sorte que $0 < \psi_{ijk} < (1 - \pi_{11}^{t1})^{-1}$.

Il convient de souligner que ces ψ_{ijk}^{ts} sont identiquement distribués sur j et sur k . De nouveau, nous pourrions utiliser d'autres distributions, telles que la densité lognormale tronquée, mais cela ne change pas grand-chose. Dans cette formulation, il existe une certaine information au sujet de β parce que nous supposons que les petits domaines partagent un effet commun. Donc, pour le domaine i , la densité de probabilité pour ψ_i est

$$g_i^3(\psi_i | \beta_i, \pi_i) = \prod_{j=1}^r \prod_{k=1}^c \prod_{t=1}^T \frac{\Gamma(\beta_i)}{\Gamma(\beta_i)} \frac{e^{-\beta_i \psi_{ijk}^{ts}}}{\beta_i \psi_{ijk}^{ts}} \left/ \int_0^{\infty} \frac{\Gamma(\beta_i)}{\Gamma(\beta_i)} \frac{e^{-\beta_i \psi_{ijk}^{ts}}}{\beta_i \psi_{ijk}^{ts}} d\psi_{ijk}^{ts} \right.$$

pour $0 < \psi_{ijk}^{ts} < (1 - \pi_{11}^{t1})^{-1}$. En procédant à la transformation $t_{ijk}^{ts} = \beta_i \psi_{ijk}^{ts}$, nous voyons que la constante de normalisation dans le dénominateur de chacun des facteurs qui figurent dans $g_i^3(\psi_i | \beta_i, \pi_i)$ est $G[\beta_i(1 - \pi_{11}^{t1})^{-1}]$, où $G(\cdot)$ est la fonction gamma avec paramètre d'échelle β_i . Pour que l'intégration ne dépende pas de π_{11}^{t1} , posons que $\phi_{ijk}^{ts} = (1 - \pi_{11}^{t1}) \psi_{ijk}^{ts}$ et que $\phi_i = (\phi_{i1}^{t1}, \dots, \phi_{i4}^{t1}, \phi_{i1}^{t2}, \dots, \phi_{i4}^{t2}, \dots, \phi_{i1}^{trc}, \dots, \phi_{i4}^{trc})'$. Alors

n'avons pas suivi cette direction pour élaborer nos modèles, en partie parce que nous n'utilisons pas de covariables ici ; voir Nandram et Choi (2010) pour l'utilisation de covariables et d'effets aléatoires.

L'approche décrite dans Nandram et Choi (2002a, b) est intéressante, mais elle ne s'applique pas directement au problème des tableaux de contingence $r \times c$ qui nous

occupe. Plus précisément, dans Nandram et Choi (2002a, b), un seul paramètre de centrage est nécessaire par domaine. Dans notre formulation, nous avons besoin de rc paramètres de centrage par domaine ; chacun de ces paramètres doit suivre une loi centrée sur l'unité pour permettre la détermination d'un modèle de non-réponse ignorable. Des continuités d'inégalité doivent également être incluses dans le modèle de non-réponse non ignorable. En outre, on ne peut écarter la possibilité que ces paramètres soient corrélés. La méthodologie nécessaire pour appliquer les travaux de Nandram et Choi (2002a, b) au tableau de contingence $r \times c$ n'est pas simple. Constatant ces difficultés, Nandram, Liu, Choi et Cox (2005) (avec un seul tableau supplémentaire) et Nandram, Cox et Choi (2005) (avec les trois tableaux supplémentaires) appliquent une idée plus simple, mais moins élégante que dans Nandram et Choi (2002a, b) pour procéder au centrage ; voir aussi Nandram et Choi (2005).

Essentiellement, Nandram, Cox et Choi (2005), ainsi que Nandram, Liu, Cox et Choi (2005) émettent l'hypothèse d'un modèle ignorable, obtennent des échantillons des probabilités de réponse et se servent de ces probabilités de réponse d'un modèle de non-réponse non ignorable en « contrôlant » ce paramètre. Naturellement, une alternative est possible lorsque l'information sur le degré de non-ignorabilité, l'intégration d'information par rapport à priori au sujet d'un écart systématique par rapport à l'ignorabilité est plus complexe dans le cas de notre problème et elle nécessiterait du travail sur le terrain supplémentaire coûteux afin d'obtenir cette information.

Nous discutons maintenant de notre conception du problème de non-réponse non ignorable, fondamentalement un problème de distorsion. En fait, ce problème est extrêmement difficile et nous devons essayer d'en trouver une solution, mais que nous devons essayer d'en trouver une. Sans aucune information, il n'est pas possible de dire quelles sont les différences entre les répondants et les non-répondants. Un modèle de non-réponse ignorable est restreint, parce qu'il suppose que les répondants et les non-répondants sont semblables, alors qu'ils peuvent être différents. Les statisticiens doivent non seulement faire face à l'imprécision (erreur d'échantillonnage), mais aussi être suffisamment audacieux pour étudier la subjectivité (l'ignorance découlant de l'information manquante).

2. Le modèle de non-réponse non ignorable

Malheureusement, il est bien connu que les modèles de non-réponse non ignorable contiennent des paramètres non identifiables. Nous discutons de la façon dont sont identifiés les paramètres clés de non-ignorabilité. Nous savons que si les répondants et les non-répondants sont semblables, les γ_j sont égaux à l'unité, et nous obtenons le modèle de non-réponse ignorable à l'aide de tous les paramètres identifiés. Nous pouvons maintenant étendre le modèle de non-réponse ignorable à un modèle de non-réponse non ignorable en donnant à ces paramètres γ_j une loi centrée à 1, tout en maintenant l'identifiabilité. Un modèle de non-réponse non ignorable peut être formulé pour ajouter de la souplesse au modèle de non-réponse ignorable, comme nous l'avons fait dans nos travaux ; la souplesse est une forme d'analyse de sensibilité, cohérente dans le cas qui nous occupe, et il s'agit en effet d'une évaluation bayésienne de l'incertitude (du risque) (par exemple Greenland 2009). C'est ce que nous avons fait ou essayé de faire dans nos travaux.

Dans le présent article, nous tentons de résoudre le problème difficile de Nandram et Choi (2002a, b) sous sa forme originale pour les tableaux $r \times c$ pour de nombreux domaines. Le plan de l'exposé est le suivant. À la section 2, nous décrivons le modèle hiérarchique bayésien. Précisément, nous décrivons le mécanisme de non-réponse non ignorable et nous construisons une loi a priori propre. À la section 3, nous montrons comment ajuster le modèle en utilisant l'algorithme d'échantillonnage avec rééchantillonnage par importance (SIR, pour *sampling importance resampling*) pour effectuer un sous-échantillonnage à partir d'une densité de probabilité a posteriori approximative après une agrégation innovatrice de la densité a posteriori conjointe complète. À la section 4, nous illustrons notre méthodologie en nous servant de données à grande diffusion reçues dans 13 États dans le cadre de la troisième édition de la National Health and Nutrition Examination Survey (NHANES III). À la section 5, nous présentons nos conclusions.

Dans le contexte du problème de non-réponse dans un tableau à double entrée, on peut avoir affaire aussi bien à la non-réponse partielle qu'à la non-réponse totale. Donc, on peut considérer la totalité du tableau de données comme étant constitué de quatre tableaux, à savoir un tableau pour les données complètes et trois tableaux supplémentaires, pour l'information de ligne manquante, pour l'information de colonne manquante et pour l'information de ligne ainsi que de colonne manquante, respectivement. Dans tous l'exposé, nous donnons aux lignes l'indice $j = 1, \dots, r$, aux colonnes l'indice $k = 1, \dots, c$, et aux quatre tableaux l'indice $s = 1, 2, 3, 4$. Nous donnons aux domaines l'indice

non ignorables. La difficulté générale que pose un modèle de non-réponse non ignorable tient au fait que les paramètres ne sont pas identifiables [par exemple, voir Nandram et Choi (2004, 2005, 2008, 2010), et Nandram, Han et Choi

Pour un tableau de contingence $r \times c$, soit I^{ijk} l'individu dans le i^e domaine se trouvant dans la j^e ligne et la k^e colonne, et 0 autrement. En outre, soit $J^n = I$ si le i^e individu dans le i^e domaine a fourni des renseignements complets et 0 autrement. Enfin, soit $P(J^n) = I \mid I^{ijn} = 1, 0, \dots, i^r, j^c, k^c \neq k) = \pi^{ijk}$. Pour la non-réponse totale (n) ou non-réponse d'une unité), si $\pi^{ijk} = \pi_i$, le modèle est ignorable ; pour la non-réponse partielle (non-réponse à certaines questions), si les valeurs de colonne manquent, mais que les valeurs de ligne sont observées et que $\pi^{ijk} = \pi_j$ (no $\pi^{ijk} = \pi_i$), le modèle est ignorable ; si les valeurs des lignes manquent, mais que les valeurs de colonne sont observées et que $\pi^{ijk} = \pi_k$ (no $\pi^{ijk} = \pi_i$), le modèle est ignorable. Tous les autres modèles sont non ignorables : voir Rubin (1976) pour une explication plus

Nandram et Choi (2002a, b) se servent d'un modèle d'extension pour étudier les données binaires en présence de non-réponse non ignorable. Le modèle d'extension, qui est un modèle de non-réponse non ignorable, dégénère en un modèle de non-réponse ignorable (dans l'esprit de Draper 1995) quand la valeur d'un paramètre de contrainte est fixée à l'unité. Cela permet d'exprimer la certitude au sujet de l'ignorabilité ; voir également Forster et Smith (1998).

Nous discutons du modèle proposé par Nandram et Choi (2002a, b) pour des données binaires provenant de petits domaines. De sorte que J^n désigne les indicateurs de réponse et I^n , la réponse binaire. Spécifiquement, en introduisant les paramètres de centrage γ_i pour le domaine i afin d'intégrer l'incertitude au sujet de l'ignorabilité, le modèle de Nandram et Choi (2002a, b) est

$$\begin{aligned} I^n &\sim \text{Bernoulli}(p), \\ J^n &\sim \text{Bernoulli}(\tau_j), j = 1, \dots, l, \\ J^n | \{\tau_j, j = 1, \dots, l\} &\sim \text{Bernoulli}(\gamma_j \tau_j), \gamma_j > 0, \end{aligned}$$

Quand $\gamma_i = 1$, le modèle de non-réponse non ignorable dégénère en un modèle de non-réponse ignorable. Ici, γ_i est le ratio des chances de succès parmi les répondants aux chances de succès parmi l'ensemble des individus pour le i^{e} domaine. Le paramètre γ_i décrit la portée de la non-ignorabilité du mécanisme de réponse pour le domaine i , et c'est donc grâce à ce paramètre γ_i qu'est intégrée l'incertitude au sujet de l'ignorabilité. Nandram et Choi (2002a, b) définissent $\delta_i = \pi_i \{ \gamma_i p_i + (1 - p_i) \}$ comme étant la probabilité

Nandram et Katzoff : Un modèle hiérarchique bayésien de non-réponse

et π_i , respectivement. Ici, les paramètres ne sont pas identifiables. Cependant, quand $\gamma_i = 1$, ils le sont tous. Autrement dit, le caractère identifiable des paramètres dépend de γ_i . Notons que si $\gamma_i = 1$, nous obtenons un modèle ignorable pour un mécanisme MAR. Comme les paramètres de ce modèle sont identifiables, il est relativement logique de l'utiliser (ou des modèles similaires) comme modèle de référence. Il faut toutefois souligner que ce modèle n'est toujours pas justifié, parce qu'il repose sur l'hypothèse que les données manquantes ressemblent aux données observées. Donc, pour rendre ce modèle de non-réponse ignorable plus souple, nous utilisons le paramètre γ_i .

Soit γ_i^{av} le nombre d'individus pour lesquels $I_i = n$, $I_i^j = n$, $v = 0, 1$ dans le i^e domaine. Alors, sous le

$$\prod_i (y_i^{100}, y_i^{101}, y_i^{110}, y_i^{111}) | \pi_i, p_i, \gamma_i \rangle \sim \text{Multinomial}\{n_i, (1 - d_i)(1 - \pi_i), (1 - d_i)\pi_i, (1 - \gamma_i)\pi_i, d_i(1 - \gamma_i)\pi_i, d_i\gamma_i\}$$

avec indépendance sur les domaines. Ici, γ^{10} et γ^{11} seulement sont observés et, par conséquent, tous les paramètres obtenus la fonction de vraisemblance γ^i de la même manière pour le tableau de contingence $r \times c$ plus complet avec données manquantes.

Nous partons d'une loi gamma et, pour permettre le centrage sur le modèle de non-réponse ignorable, nous devons sélectionner chaque γ_i de manière que sa moyenne soit égale à 1. Cependant, nous devons utiliser une loi gamma tronquée, parce que $0 < \pi_i < 1$ et $0 < \gamma_i \leq 1/\pi_i$. Une idée intéressante de Nandram et Choi (2002a,b) consiste à modéliser le centrage sous forme d'une loi

$$\text{ii)} \quad \gamma_i^l \mid v \sim \text{Gamma}(v, v), 0 < \gamma_i^l < 1/\pi_i, 0 < \pi_i < 1.$$

Le modèle est complet et possède des densités de probabilité a priori non informatives sur tous les hyperparamètres. D'autres distributions peuvent être choisies (par exemple, la densité lognormale tronquée) pour les γ_i , mais il ne s'agit pas d'un problème essentiel et cela n'aurait pas beaucoup d'importance.

On peut se servir d'un modèle au niveau du domaine avec effets aléatoires dans lequel, conditionnellement aux données observées, la non-réponse dépend des effets aléatoires au niveau du domaine. Ce modèle peut être formulé en utilisant une fonction de lien logit, mais nous

Un modèle hiérarchique bayésien de non-réponse pour les données catégoriques d'un tableau à double entrée provenant de petits domaines avec incertitude au sujet de l'ignorabilité

Balagobin Nandram et Myron Katzoff¹

Résumé

Nous étudions le problème de la non-réponse non ignorable dans un tableau de contingence bidimensionnel qui peut être créé individuellement pour plusieurs petits domaines en présence de non-réponse partielle ainsi que totale. En général, le fait de prendre en considération les deux types de non-réponse dans les données sur les petits domaines accroît considérablement la complexité de l'estimation des paramètres du modèle. Dans le présent article, nous conceptualisons le tableau complet des données pour chaque domaine comme étant constitué d'un tableau contenant les données complètes et de trois tableaux supplémentaires pour les données de ligne manquantes, les données de colonne manquantes et les données de ligne et de colonne manquantes, respectivement. Dans des conditions de non-réponse non ignorable, les probabilités de cellule peuvent varier en fonction du domaine, de la cellule et de ces trois types de « données manquantes ». Les probabilités de cellule sous-jacentes (c'est-à-dire celles qui s'appliqueraient s'il était toujours possible d'obtenir une classification complète) sont produites pour chaque domaine à partir d'une loi commune et leur similitude entre les domaines est quantifiée paramétriquement. Notre approche est une extension de l'approche de sélection sous non-réponse non ignorable étudiée par Nandram et Choi (2002a, b) pour les données binaires; cette extension crée une complexité supplémentaire qui découle de la nature multivariée des données et de la structure des petits domaines. Comme dans les travaux antérieurs, nous utilisons un modèle d'extension centré sur un modèle de non-réponse ignorable de sorte que la probabilité totale de cellule dépend de la catégorie qui représente la réponse. Notre étude s'appuie sur des modèles hiérarchiques bayésiens et des méthodes Monte Carlo par chaîne de Markov pour l'inférence a posteriori. Nous nous servons de données provenant de la troisième édition de la National Health and Nutrition Examination Survey pour illustrer les modèles et les méthodes.

Mots-clés : Échantillonnage de Metropolis-Hastings ; algorithme SIR ; modèle de non-réponse non ignorable ; modèle d'extension.

1. Introduction

Généralement, les données des enquêtes par sondage sont résumées dans des tableaux de contingence à double entrée. Nous considérons le problème de la non-réponse non ignorable pour un grand nombre de tableaux de contingence de dimensions $r \times c$, pour chacun des domaines spécifiques. Dans nombre de ces enquêtes, des données manquent, si bien que la classification des individus échantillonnés n'est que partielle. Chaque tableau à double entrée présente donc à la fois des cas de non-réponse partielle (données manquantes pour l'une des deux catégories) et des cas de non-réponse totale (données manquantes pour les deux catégories). Comme on ne sait pas nécessairement de quelle façon les données manquent, il peut être souhaitable de privilégier un modèle dans lequel existe une certaine différence entre les données observées et les données manquantes (c'est-à-dire que les données manquantes ne sont pas ignorables). Pour un tableau de contingence $r \times c$ général, nous abordons la question de l'estimation des probabilités de cellule des tableaux à double entrée lorsque la non-réponse est peut-être non ignorable, mais que l'on ne dispose vraiment d'aucune information au

sujet de l'ignorabilité. Dans de telles conditions, nous aimerions exprimer le degré d'incertitude au sujet de l'ignorabilité. Nandram et Choi (2002a, b) ont décrit un modèle d'extension approprié pour les données binaires lorsqu'il existe des données provenant de nombreux petits domaines. Nous étendons ces travaux aux tableaux de contingence $r \times c$.

En désignant par x les covariables et par y la variable réponse, Little et Rubin (2002) décrivent trois types de mécanismes de création de données manquantes. Ils différencient selon que la probabilité de réponse a) est indépendante de x et de y ; b) dépend de x , mais non de y , ou c) dépend de y et éventuellement de x . Les données manquent entièrement au hasard (MCAR, *missing completely at random*) sous (a), manquent au hasard (MAR, *missing at random*) sous (b) et ne manquent pas au hasard (MNAR, *missing not at random*) sous (c). Les modèles pour les mécanismes de création de données manquantes MCAR et MAR sont dits ignorables si les paramètres de la variable dépendante du modèle et ceux de la variable réponse sont distincts (Rubin 1976). Les modèles pour les mécanismes de création des données manquantes de type MNAR sont dits

1. Balagobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, Courriel : balaman@wp.edu ; Myron Katzoff, Office of research and methodology, National Center for Health Statistics, CDC, 3311 Toledo Road, Hyattsville, MD 20782, Courriel : mjk5@cdc.gov.

Drechsler, J., et Reiter, J.P. (2010). Sampling with synthesis: A new *American Statistical Association*, 105, 1347-1357.

Elliot, M., et Purdam, K. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records. *Environment and Planning A*, 39, 1101-1118.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Reiter, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 203-211.

Reiter, J.P. (2004). Utilisation simultanée de l'imputation multiple pour les données manquantes et le contrôle de la divulgation. *Techniques d'enquête*, 30, 263-271.

Reiter, J.P. (2005). Estimating identification risks in microdata. *Journal of the American Statistical Association*, 100, 1103-1113.

Reiter, J.P. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*, 95, 933-946.

Wiltenborg, L., et de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York : Springer-Verlag.

Winkler, W.E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Rapport technique, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Remerciements

Ces travaux ont été financés par une bourse du U.S. National Science Foundation (SES-0751671).

Bibliographie

Drechsler, J., et Reiter, J.P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. Dans *Privacy in Statistical Databases*, (Eds., J. Domingo-Ferrer et Y. Saygin), New York : Springer, 227-238.

Drechsler, J., et Reiter, J.P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *Journal of Official Statistics*, 25, 589-603.

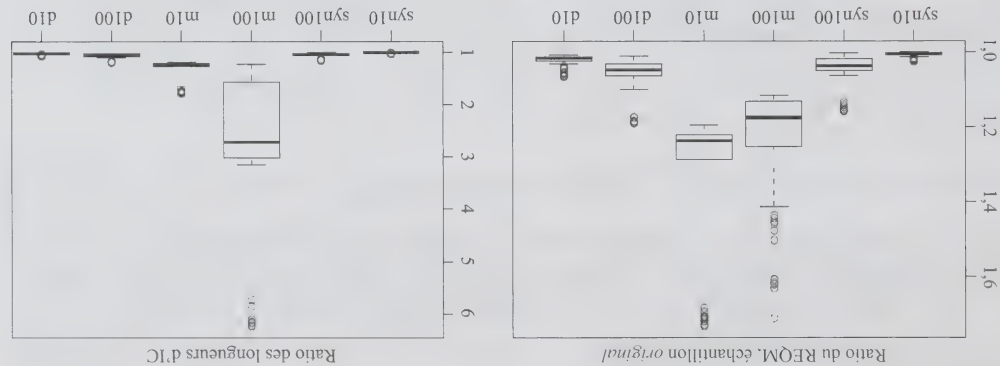


Figure 2 Comparaisons d'efficacité pour le plan d'échantillonnage stratifié. Dans les étiquettes, *original* et *syn* font référence à l'échantillon original et à l'échantillon synthétique avant le sous-échantillonnage, et *m*, *d* et les chiffres ont la même signification qu'à la figure 1. Les dénominateurs de la REQM sont basés sur les estimations ponctuelles d'après l'échantillon original sans synthèse. Chaque boîte à moustaches englobe 50 paramètres estimés

4. Conclusion

L'approche des sous-échantillons différents et celle des mêmes sous-échantillons ont des avantages concurrents. Pour un nombre fixe d'ensembles de données diffusés M , l'approche des sous-échantillons différents permet une estimation plus efficace que l'approche des mêmes sous-échantillons, comme en témoigne la figure 2, puisque les sous-échantillons diffusés sont indépendants plutôt que corrélés. L'approche des sous-échantillons différents garantit aussi des estimations de variance positives, contrairement à l'approche fondée sur le même échantillon. Cependant, si M est grand, l'approche des sous-échantillons différents affaiblit la protection de la confidentialité découlant du sous-échantillonnage, puisque les ensembles de données combinés contiennent vraisemblablement la plupart des enregistrements issus de l'enquête originale. Donc, à moins que le taux de sous-échantillonnage soit faible (par exemple 1 % ou 2 %), l'INS peut être obligé de choisir une valeur de m modeste (par exemple, $m = 5$) pour utiliser l'approche des sous-échantillons différents. Par conséquent, cette approche n'est pas viable quand la taille de l'échantillon original est modeste.

Au lieu de procéder au sous-échantillonnage avec synthèse, les organismes statistiques pourraient diffuser des données partiellement synthétiques qui comprennent tous les enregistrements provenant de l'échantillon original, pour autant qu'ils soient prêts à diffuser des fichiers de cette taille. La synthèse partielle sur les données originales produit généralement des estimations dont la variance est plus faible que le sous-échantillonnage avec synthèse, comme le

montre la figure 2, puisque qu'un plus grand nombre d'enregistrements sont diffusés. Cependant, la synthèse partielle sur les données originales engendre généralement de plus grands risques de divulgation que le sous-échantillonnage avec synthèse, puisqu'un plus grand nombre d'enregistrements à risque figurent parmi les données diffusées et que la protection supplémentaire due au sous-échantillonnage fait défaut. Les organismes statistiques peuvent comparer les deux options en fonction des risques de divulgation en se servant des méthodes de Drechsler et Reiter (2008), qui tiennent compte de la protection résultant de l'échantillonnage, ainsi qu'en fonction de l'utilité des données en comparant les inférences pour les analyses représentatives. Il se peut également que le processus de sous-échantillonnage engendre une protection supplémentaire suffisante pour permettre une quantité plus faible de synthèse qu'il ne serait nécessaire dans le cas d'une synthèse partielle de l'ensemble de données originales complètes. L'évaluation de l'utilité des données pour le sous-échantillonnage avec synthèse par opposition à la synthèse seulement pour des risques de divulgation donnés dépasse le cadre de ce bref exposé, mais il s'agit d'un domaine intéressant dans lequel poursuivre de futures études.

Nous n'avons pas élaboré d'approches de sous-échantillonnage avec synthèse pour d'autres plans de sondage que l'échantillonnage aléatoire simple (stratifié). Pour l'approfondissement des sous-échantillons différents, les méthodes d'inférence appropriées nécessitent une estimation approximativement sans biais de la variance provenant de la première phase d'échantillonnage qui peut être calculée d'après le sous-échantillon seulement. Sous des plans de sondage

Le graphique de droite de la figure 1 donne les ratios de la racine carrée de l'erreur quadratique moyenne (REQM) simulée de \bar{q}_m sur les REQM simulées pour les sous-échantillons sans aucune synthèse. Pour l'approche fondée sur le même sous-échantillon, les REQM des sous-échantillons synthétiques ont tendance à être plus faibles que celles calculées pour les sous-échantillons sans aucune synthèse, particulièrement dans le cas de la synthèse totale. Les REQM plus faibles sont dues au fait que les modèles synthétiques sont déterminés en se servant de D , c'est-à-dire les données d'enquête avant de tirer le sous-échantillon, de sorte qu'ils contiennent des renseignements supplémentaires qui ne figurent pas dans les sous-échantillons sans synthèse. Pour les divers sous-échantillons synthétiques, les ratios des REQM sont habituellement supérieurs à 1. Ici, l'accroissement de la synthèse entraîne une plus grande perte d'efficacité. Il convient de souligner qu'à la figure 1, les REQM pour l'approche fondée sur différents échantillons et celle fondée sur le même échantillon ne sont pas directement comparables, parce qu'elles sont basées sur des dénominateurs différents.

Afin de permettre de comparer les méthodes, et pour illustrer les pertes d'efficacité dues au sous-échantillonnage, nous répétons le plan de simulation en utilisant $m = 25$ pour l'approche des sous-échantillons indépendants et échantillonage.

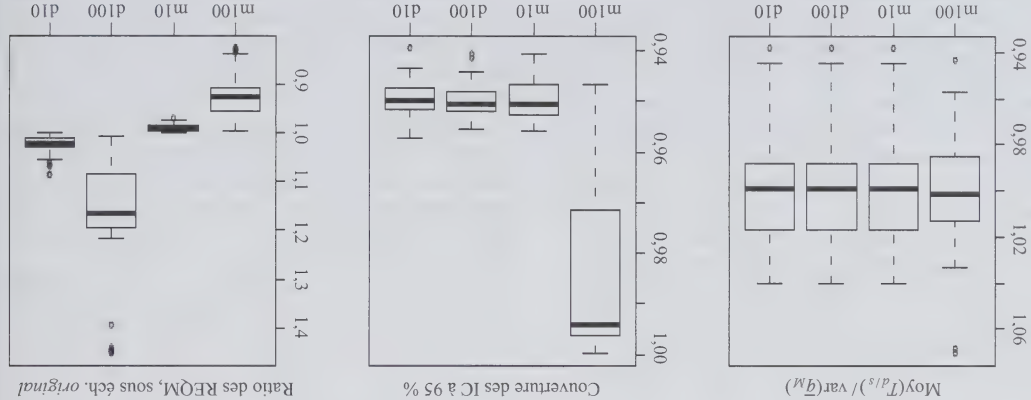


Figure 1 Résultats des simulations pour le plan d'échantillonnage stratifié. Dans les étiquettes, m et d indiquent l'approche fondée sur le même sous-échantillon et celle fondée sur des sous-échantillons différents. Les chiffres donnent le pourcentage d'enregistrements qui sont synthétisés. Les dénominateurs de la REQM sont fondés sur les estimations ponctuelles d'après les sous-échantillons sans synthèse. Pour l'approche des sous-échantillons différents, la REQM est calculée en se servant de la moyenne des m estimations ponctuelles. Chaque boîte à moustaches comprend 50 paramètres estimés

correspondante pour les 50 paramètres estimés. Les ratios médians sont proches de 1 dans tous les scénarios, et les moyennes de T_d (et de T_s) ne diffèrent jamais de plus de 10 % des variances réelles correspondantes. Donc, T_s et T_d semblent être tous deux des estimateurs de variance app-
proximativement valides.
Le graphique du milieu de la figure 1 résume les pour-
centages des 5 000 intervalles de confiance à 95 % synthé-
tiques basés sur T_d (et sur T_s) qui couvrent leur grandeur
 \hat{Q} correspondante. Les taux de couverture sont proches de
0,95, excepté pour les coefficients de régression sous l'ap-
proche fondée sur le même sous-échantillon avec synthèse
totale. Pour ces coefficients, $T_s < 0$ dans une proportion
allant jusqu'à 38 % des exécutions de la simulation, de sorte
que les intervalles de confiance sont fondés sur l'estimateur
 T_s^* prudent. La fraction plus élevée de variances négatives
survient dans la strate la plus petite dont le taux d'échantil-
lonnage est de 50 %. Toutes les estimations de variance sont
positives quand 10 % seulement des enregistrements sont
synthétisés.

Dans le cas de l'approche fondée sur des sous-échan-
tillons différents, nous générerons $m = 5$ enquêtes synthé-
tiques comme il est décrit à la section 2.1. Pour l'approche
fondée sur le même sous-échantillon, nous commençons par
tirer $m = 5$ valeurs de θ , les coefficients de régression et
les variances. Pour chaque $\theta^{(i)}$, nous générerons $r = 5$ en-
sembles de données synthétiques pour chaque emboîtement
(*nest*) de premier degré.
Pour tous les scénarios, nous répétons le processus
consistant à i) créer D par échantillonnage à partir de la
population et ii) générer des sous-échantillons avec synthèse
5 000 fois en tout. Pour chacune de ces 5 000 exécutions,
nous obtenons des inférences pour 50 paramètres, y compris
les moyennes de population et les moyennes de strates de
 X_4 et X_5 , les coefficients d'une régression de X_3 sur toutes
 X_5 sur toutes les autres variables. Les régressions sont esti-
mées séparément dans chaque strate.
La figure 1 donne les principaux résultats des simula-
tions. Le graphique de gauche présente les ratios de la
moyenne simulée de T_d (et de T_s) à la var(\hat{Q}_m) simulée

Tableau 2
Paramètres pour le tirage de (X_2, \dots, X_5) pour la population

	Taille de strate	Modèle	Distribution du terme d'erreur
Strate 1	750 000	$X_2 = X_1 + e$ $X_3 = X_1 + X_2 + e$ $X_4 = X_1 + X_2 + X_3 + e$ $X_5 = X_1 + X_2 + X_3 + X_4 + e$	$e \sim N(0, 5)$
Strate 2	200 000	$X_2 = 2X_1 + e$ $X_3 = 2X_1 + 0,5X_2 + e$ $X_4 = 2X_1 + 0,5X_2 + X_3 + e$ $X_5 = 2X_1 + 0,5X_2 + X_3 + e$	$e \sim N(0, 10)$
Strate 3	40 000	$X_2 = -3X_1 + e$ $X_3 = -3X_1 - 1,5X_2 + e$ $X_4 = -3X_1 + X_2 - 1/3X_3 + e$ $X_5 = -3X_1 + X_2 - 1/3X_3 + 1/9X_4 + e$	$e \sim N(0, 30)$
Strate 4	10 000	$X_2 = -2X_1 + e$ $X_3 = -X_1 - 1,5X_2 + e$ $X_4 = -2X_1 + X_2 + 1/4X_3 + e$ $X_5 = 2X_1 - X_2 - 1/4X_3 + 1/16X_4 + e$	$e \sim N(0, 20)$

Il est dérivé par appariement des moments à ceux d'une variable aléatoire χ^2 .

3. Simulations illustratives sous un plan d'échantillonnage stratifié

À la présente section, nous examinons les propriétés analytiques des procédures d'inférence proposées pour le sous-échantillonnage avec synthèses dans le cas de l'échantillonnage aléatoire simple stratifié. Nous générons une population de $N = 1\,000\,000$ enregistrements comprenant cinq variables, X_1, \dots, X_5 , dans $H = 4$ strates. X_1 est une variable catégorique comprenant dix catégories générales conformément à la distribution du tableau 1. Les distributions pour (X_2, \dots, X_5) sont présentées au tableau 2, de même que les tailles de strates.

Tableau 1
Distribution empirique de X_1 dans la population générale

	1	2	3	4	5	6	7	8	9	10
Pourcentage	24,77	32,63	16,38	15,06	7,13	2,53	0,95	0,33	0,15	0,09

Pour créer D , nous tirons aléatoirement $n_h = 7\,500$ enregistrements dans chaque strate. Chaque sous-échantillon comprend $n_h = 5\,000$ enregistrements pour chaque strate.

En pratique, l'INS pourrait recourir à la répartition proportionnelle pour fixer chaque n_h et choisir des taux d'échantillonnage plus faibles pour fixer n_h . Nous utilisons une taille d'échantillon commune et de grandes fractions de variance pour le sous-échantillonnage avec synthèses traite correctement les facteurs de correction pour population finie non négligeables, par exemple 50 % des enregistrements sont échantillonnés dans la strate 4.

Nous considérons que Y_4 et Y_5 sont des variables confidentielles et illustrons deux scénarios de synthèse. Dans le premier, nous synthétisons les valeurs de Y_4 et Y_5 de tous les enregistrements. Pour cela, dans chaque strate, nous simulons Y_{4h} en utilisant une régression de X_{4h} sur (X_{1h}, X_{2h}, X_{3h}) estimée en nous servant de D , et nous simulons Y_{5h} en utilisant une régression de X_{5h} sur $(X_{1h}, X_{2h}, X_{3h}, X_{4h})$ estimée en nous servant de D . Les prédictions de Y_{4h} sont fondées sur les valeurs synthétisées de X_{4h} . Dans la deuxième approche, dans chaque strate, nous ne remplaçons que dans la strate h . Nous générons des valeurs de population dans la strate h ou $p_h > p_h^*$, où p_h^* est le 90^e centile de X_3 dans la population. Les modèles en nous servant uniquement des enregistrements qui satisfont $X_{3h} > p_h^*$.

$d^* = \{d^{(l,p)} : l = 1, \dots, m; p = 1, \dots, r\}$. Chaque $d^{(l,p)}$ comprend un indice l désignant son emboîtement (*nest*). Pour $l = 1, \dots, m$ et $p = 1, \dots, r$, soit $q^{(l,p)}$ et $n^{(l,p)}$ l'estimation de \bar{Q} et l'estimation de sa variance calculée en se servant de $d^{(l,p)}$. Ici, $n^{(l,p)}$ comprend le facteur de correction pour population finie. Les quantités qui suivent sont utilisées pour les inférences :

$$\bar{q} = \sum_{m=1}^m \sum_{r=1}^r b^{(l,p)} / (mr), \quad (7)$$

$$\bar{w}^M = \sum_{m=1}^m \sum_{r=1}^r (b^{(l,p)} q^{(l,p)})^2 / (m - 1), \quad (8)$$

$$b^M = \sum_{m=1}^m (b^{(l,p)})^2 / (m - 1), \quad (9)$$

$$\bar{n} = \sum_{m=1}^m \sum_{r=1}^r n^{(l,p)} / (mr). \quad (10)$$

L'analyste peut se servir de \bar{q} pour estimer \bar{Q} et de $T = \bar{n} - \bar{w}^M + (1 + 1/m) b^M - \bar{w}^M / r$ pour estimer la variance de \bar{q}^M . Quand r est grand, les inférences sont fondées sur une loi $t_r(\bar{q}^M - \bar{Q}) \sim t_{r_s}(0, T_s)$, dont le nombre de degrés de liberté est

$$v_s = \left(\frac{(1 + 1/m) b^M q^{(l,p)}}{(1 + 1/r) \bar{w}^M} \right)^2 + \frac{(m - 1) T_s^2}{\{m(r - 1)\} T_s^2}. \quad (11)$$

Il se peut que $T_s < 0$, particulièrement si m et r sont petits. L'analyste peut utiliser à la place l'estimateur de variance toujours positif, mais prudent, $T_s^* = \lambda T_s + (1 - \lambda)(1 + 1/m) b^M q^{(l,p)}$, où $\lambda = 1$ quand $T_s > 0$ et $\lambda = 0$, autrement. Les facteurs qui motivent l'emploi de cet estimateur sont donnés dans Reiter (2008). En général, il est possible d'éviter les valeurs négatives de T_s en donnant une valeur élevée à m et à r . Quand $T_s < 0$, les inférences sont fondées sur une loi t à $(m - 1)$ degrés de liberté qui découle de l'utilisation du premier terme seulement et de T_s^* dans (11). Pour les plans stratifiés, l'estimation ponctuelle des quantités pour l'ensemble de la population est $\bar{q}^M = \sum_h (N_h / N) \bar{q}^{Mh}$, et sa variance estimée est $T_s = \sum_h (N_h / N)^2 T_{sh}$, où \bar{q}^{Mh} et T_{sh} sont l'estimation ponctuelle et sa variance dans la strate h . Le nombre de degrés de liberté de la loi t pour l'échantillonnage stratifié est égal à

$$v_s = \left(\frac{\sum_h (N_h / N)^2 (1 + 1/m) b^{Mh}}{\sum_h (N_h / N)^2 (1 + 1/r) \bar{w}^{Mh}} \right)^2 + \frac{\{m(r - 1)\} T_s^2}{\sum_h (N_h / N)^2 T_{sh}^2}. \quad (12)$$

de la moyenne de la population totale, l'estimation ponctuelle de \bar{Q} est $\bar{q}_m = \sum h_i (N_i / N) \bar{q}_{mh_i}$, et sa variance estimée est $T_d = \sum h_i (N_i / N)^2 T_{dih_i}$. Les estimations ponctuelles et les estimations de variance pour des fonctions non linéaires des moyennes peuvent être obtenues en utilisant des développements en série de Taylor. Il convient de mentionner que les INS devraient diffuser les valeurs de n_{2h} / n_{10} pour toutes les strates pour permettre l'estimation

2.1.2 Dérivation des inférences pour l'approche des sous-échantillons aléatoires différents

L'analyste s'intéresse à $f(\bar{Q} | d^{syn})$, qui peut s'écrire sous la forme

$$f(\bar{Q} | d^{syn}) = \int f(\bar{Q} | d^{syn}) f(d^{syn} | d^{syn}) d d^{syn}. \quad (5)$$

Dans toutes les dérivations mathématiques de la section 2.1.2, nous supposons que l'analyste utilise les mêmes distributions que celles employées par l'INS pour créer D^{syn} . Nous supposons aussi que les tailles d'échantillons sont suffisantes pour permettre des approximations normales de ces distributions. Donc, pour chaque distribution, nous n'avons besoin que des deux premiers moments, que nous dérivons en nous servant d'arguments bayésiens en grand échantillon classiques. Nous émettons l'hypothèse de

Soit \bar{Q}_i et U_i l'estimation ponctuelle de \bar{Q} et sa variance que l'analyste calculerait en se servant de D_i (dont il ne dispose pas). Soit les quantités $\bar{Q}_m | U_m$, et B_m définies dans (1) à (3), mais en utilisant \bar{Q}_i et U_i . Partant des résultats de la synthèse par la classe (Reiter 2003), nous avons $(\bar{Q} | D^{syn}) \sim N(\bar{Q}_m, \bar{U}_m + B_m / m)$. Nous supposons que $(q_i | D_i) \sim N(\bar{Q}_i, (1 - n_i / n_1) u_i)$ et, comme il est habituel dans les contextes d'imputation multiple, que $u_i \approx \bar{u}_m$. Donc, en utilisant la théorie bayésienne classique, nous avons $(\bar{Q}_m | d^{syn}) \sim N(\bar{Q}_m, \bar{U}_m + B_m / m + (1 - n_i / n_1) \bar{u}_m / m)$. Pour obtenir $f(\bar{Q} | d^{syn})$, nous devons éliminer par intégration B_m et \bar{U}_m de cette distribution. Pour cela, nous substituons à B_m et à \bar{U}_m leurs valeurs prévues approximatives. Pour approximer $E(B_m | d^{syn})$, nous utilisons $b_m = (1 - n_2 / n_1) \bar{u}_m$. Pour approximer $E(\bar{U}_m | d^{syn})$, nous obtenons

$$\begin{aligned} \text{Var}(\bar{Q} | d_i) &= E[\text{var}(\bar{Q} | D_i) | d_i] + \text{var}[E(\bar{Q} | D_i) | d_i] \\ &= E(U_i | d_i) + \text{var}(\bar{Q}_i | d_i) \\ \text{Ici, } \text{var}(\bar{Q} | d_i) &= (1 - n_2 / N) u_i. \text{ En résolvant (6), nous obtenons } E(\bar{U}_m | d^{syn}) \approx (n_2 / n_1 - n_2 / N) \bar{u}_m. \text{ Après substitution de ces valeurs prévues, nous avons } \text{var}(\bar{Q} | d^{syn}) = T_d. \end{aligned} \quad (6)$$

Puisque nous utilisons une variance estimée pour \bar{Q} , nous approximations $f(\bar{Q} | d^{syn})$ par une loi t de moyenne \bar{q}_m et de variance T_d . Le nombre de degrés de liberté, ν_d , est dérivé en faisant concorder les deux premiers moments de $(\nu_d T_d) / (n_2 / n_1 - n_2 / N) \bar{u}_m + B_m / m + (1 - n_2 / n_1) \bar{u}_m / m$ avec ceux d'une loi $\chi^2_{\nu_d}$.

2.2 Diffusion du même sous-échantillon aléatoire

À première vue, diffuser un même ensemble d'enregistrements dans chaque sous-échantillon ressemble à la synthèse par la classe. Cependant, l'estimateur de variance de Reiter (2003) peut présenter un biais positif dans ce contexte. Pour l'illustrer, supposons que D ne contient qu'une seule variable dont la moyenne d'échantillon est \bar{x}_1 . Supposons aussi que nous créons D^{syn} en remplaçant toutes les valeurs de x_i et que nous sélectionnons aléatoirement un ensemble commun de n_2 enregistrements pour le sous-échantillon. Soit $m = \infty$, et soit \bar{Q} la moyenne de la population de x . Si les remplacements sont simulés d'après le modèle correct, qui est estimé en se servant de D , alors $\bar{q}_\infty = \bar{x}_1$. D'où, $\text{var}(\bar{q}_\infty)$ est identique à la variance de \bar{x}_1 , qui est $(1 - n_1 / N) s_1^2 / n_1$. Toutefois, l'estimation de variance de Reiter (2003) comprend \bar{u}_m basé sur $(1 - n_2 / N) s_2^2 / n_2$ ou $E(s_2^2) = s_1^2$. Donc, en général, la variance de Reiter (2003) aura un biais positif pour les sous-échantillons avec

Au lieu de la synthèse par la classe, nous adoptons l'approche suivie par Reiter (2008) sous imputation multiple pour remplacer les données manquantes quand les enregistrements qui servent pour l'imputation ne sont ni utilisés ni diffusés pour l'analyse. Ces conditions ressemblent au tirage des mêmes enregistrements dans chaque sous-échantillon d_i parce que les modèles utilisés pour la synthèse sont estimés en se servant de D , mais l'analyste ne dispose que de d^{syn} pour l'analyse; autrement dit, les enregistrements utilisés pour l'imputation ne sont pas tous diffusés pour l'analyse. Par souci de commodité, nous résumons ici la méthodologie de Reiter (2008) sans inclure les dérivations mathématiques. Premièrement, comme dans la synthèse par la classe, l'INS estime les modèles de synthèse dont les valeurs seront uniquement les enregistrements dont les valeurs seront synthétisées. Soit θ les paramètres qui régissent la distribution des modèles de données synthétiques. Deuxièmement, l'INS tire un échantillon de m valeurs de θ de la distribution tirage $\theta^{(l)}$ où $l = 1, \dots, m$. L'INS tire un ensemble de données de remplacement $D^{(l,p)}$ des modèles de synthèse fondés sur $\theta^{(l)}$. L'INS répète ce processus r fois pour chaque $\theta^{(l)}$. Enfin, l'INS diffuse la série de $M = mr$ sous-échantillons provenant de ces ensembles de données,

sensibles présentes dans D par des imputations multiples. Les modèles de synthèse sont estimés en ne se servant que des enregistrements dont les valeurs seront synthétisées. La synthèse D^{syn} est effectuée indépendamment m fois, ce qui donne D^{syn} . L'INS tire alors un sous-échantillon aléatoire simple de $n_2 < n_1$ enregistrements de chaque D_i . Ces m sous-échantillons, $d_i^{syn} = \{d_i^j : i = 1, \dots, m\}$, sont diffusés aux membres du public.

L'analyste de d^{syn} cherche à faire des inférences au sujet d'un paramètre \bar{Q} , par exemple une moyenne de population ou un coefficient de régression. Dans chaque sous-échantillon d_i , il estime \bar{Q} au moyen d'un estimateur ponctuel q_i , et la variance de q_i , au moyen d'un estimateur u_i , en spécifiant q et u comme si d_i comprenait les données collectées durant l'enquête. Ici, u est spécifié en faisant abstraction de tout facteur de correction pour population finie ; par exemple, quand q est la moyenne d'échantillon, $u = s^2/n_2$, où s^2 représente la variance de q et u dans d_i . Les quantités

$$(1) \quad \bar{q}^w = \sum_{i=1}^m \bar{q}_i^w / m$$

$$(2) \quad b^w = \sum_{i=1}^m (q_i^w - \bar{q}^w)^2 / (m - 1)$$

$$(3) \quad \bar{u}^w = \sum_{i=1}^m u_i^w / m.$$

L'analyste peut alors utiliser \bar{q}^w pour estimer \bar{Q} et

$$(4) \quad T^d = (n_2/n_1 - n_2/N) \bar{u}^w + b^w/m$$

pour estimer la variance de \bar{q}^w . Les dérivations de ces estimations sont présentées à la section 2.1.2. Il convient de noter que, sans le sous-échantillonnage, c'est-à-dire si $n_2 = n_1$, (4) est égal à l'estimation de la variance pour la synthèse partielle classique (Reiter 2003). Quand n_2 est grand, les inférences sont fondées sur une loi $t_{\nu}^d(0, T^d)$, dont le nombre de degrés de liberté est égal à $\nu^d = (m - 1) (1 + (n_2/n_1 - n_2/N) \bar{u}^w / b^w)^2$.

Les méthodes d'inférence peuvent être étendues à l'échantillonnage stratifié dans les conditions où l'INS utilise les mêmes strates pour le sous-échantillon et l'échantillon original. Soit N_h la taille de la population dans la strate h , où $h = 1, \dots, H$. Pour chaque h , soit \bar{q}_h^w et T_h^w les valeurs de (1) et (4) calculées en se servant uniquement des enregistrements compris dans la strate h . Ces estimations sont utilisées dans les inférences des quantités de population dans la strate h . Pour les inférences au sujet

and Program Participation, la Longitudinal Business Data-base, la Survey of Consumer Finances, les données sur les logements de groupe (*group quarters*) de l'American Community Survey et l'application Web *OnTheMap*. L'approche proposée ici diffère de la synthèse partielle en raison du sous-échantillonnage, qui nécessite que l'on ajuste les méthodes inférentielles de Reiter (2003) ; ces ajustements sont présentés ici. L'approche diffère aussi des méthodes de création d'échantillons de microdonnées à grande diffusion synthétiques d'après des données de recensement élaborées récemment par Drechsler et Reiter (2010). Dans le sous-échantillonnage avec synthèse, les données initiales proviennent d'une enquête et non d'un recensement ; donc, les inférences doivent tenir compte de l'incertitude additionnelle qui résulte de l'échantillonnage initial.

2. Approche générale

Nous décrivons maintenant la génération des données et les procédures d'inférence pour les deux approches de sous-échantillonnage avec synthèse, à savoir la diffusion d'un échantillon (indépendants) ou la diffusion d'un même ensemble d'enregistrements dans chaque sous-échantillon. Les méthodes de génération des données, ainsi que celles permettant de faire des inférences valides d'après plusieurs ensembles de données dépendent de l'approche de sous-échantillonnage. Dans les deux approches, nous désignons par D les données d'échantillon originales recueillies auprès de n_1 unités échantillonnées dans une population constituée de N unités. Nous supposons pour commencer que le plan de sondage original est un échantillonnage aléatoire simple ; par après, nous étendons l'exposé à l'échantillonnage stratifié. Nous supposons que toutes les unités échantillonnées répondent complètement dans D . Contrairement à l'approche de synthèse partielle classique (Reiter 2004), les méthodes présentées ici n'ont pas été élaborées en vue de traiter simultanément les données manquantes et la synthèse avec sous-échantillonnage. Nous nous concentrons ici sur les descriptions générales des approches et sur la présentation des méthodes inférentielles. Nous ne discutons pas des stratégies de construction de modèles de synthèse ; voir Drechsler et Reiter (2009) et les références incluses pour des directives à ce sujet.

2.1 Diffusion de différents sous-échantillons aléatoires

2.1.1 Résumé de l'approche

Pour commencer, l'INS crée m ensembles de données partiellement synthétiques, $D^{syn} = \{D_i^j : i = 1, \dots, m\}$, pour l'enquête originale en suivant l'approche de Reiter (2003). En particulier, l'INS remplace les valeurs identifiantes ou

Conjuguer des données synthétiques et le sous-échantillonnage pour créer des fichiers de microdonnées à grande diffusion pour les enquêtes à grande échelle

Jörg Drechsler et Jérôme P. Reiter ¹

Résumé

Afin de créer des fichiers de données à grande diffusion à partir d'enquêtes à grande échelle, les organismes statistiques diffusent parfois des sous-échantillons aléatoires des enregistrements originaux. Le sous-échantillonnage aléatoire amenuise la taille des fichiers transmis aux analystes secondaires des données et réduit les risques de divulgation accidentelle de renseignements confidentiels sur les participants aux enquêtes. Cependant, le sous-échantillonnage n'élimine pas entièrement le risque, de sorte qu'il faut altérer les données avant leur diffusion. Nous proposons de créer des sous-échantillons protégés contre la divulgation provenant d'enquêtes à grande échelle en recourant à l'imputation multiple. L'idée consiste à remplacer dans l'échantillon original les valeurs identificatoires ou sensibles par des valeurs tirées de modèles statistiques et de diffuser des sous-échantillons de ces données protégées contre la divulgation. Nous présentons des méthodes permettant de faire des inférences fondées sur les multiples sous-échantillons synthétiques.

Mots clés : Confidentialité ; divulgation ; imputation multiple.

1. Introduction

Les instituts nationaux de la statistique (INS) tels que le U.S. Census Bureau et Statistique Canada réalisent des enquêtes à grande échelle, comme l'American Community Survey (ACS) et l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ), qui sont jugées précieuses par les analystes secondaires des données. Alors que ceux-ci souhaitent avoir accès à autant de données que possible, l'INS doit veiller à ce que l'identité des participants à l'enquête et les attributs sensibles demeurent confidentiels. Une stratégie souvent adoptée pour réduire les risques de divulgation dans les études à grande échelle consiste à diffuser des sous-échantillons des données d'enquête originales. Ainsi, le Census Bureau diffuse un sous-échantillon des données de l'ACS englobant 1 % de l'ensemble des ménages américains (les données recueillies durant l'ACS couvrent 2,5 % de ces ménages), et Statistique Canada diffuse les données sur un échantillon de 20 % des personnes qui participent à l'ELNEJ. Pour une discussion de la protection de la confidentialité découlant de l'échantillonnage, consulter Willemboer et de Waal (2001), ainsi que Reiter (2005). Toutefois, le sous-échantillonnage n'élimine habituellement pas à lui seul les risques de divulgation, surtout si certaines unités du sous-échantillon présentent des combinaisons inhabituelles de caractéristiques. Par conséquent, les INS altèrent les données avant leur diffusion. Par exemple, dans le cas de l'ACS, le Census Bureau procède à la permutation de certaines données, au regroupement des valeurs extrêmes supérieures de certaines variables, à

l'aggrégation géographique et à la perturbation des données sur l'âge, dans le cas de l'ELNEJ, Statistique Canada recourt à la permutation et à la suppression de données. Quand elles sont appliquées intensivement, comme cela est parfois nécessaire pour assurer le respect de la confidentialité dans des enquêtes suscitant beaucoup d'attention, les stratégies classiques de contrôle de la divulgation peuvent fausser considérablement les inférences (Winkler 2007; Elliott et Purdam 2007; Drechsler et Reiter 2010). En outre, dans le cas de nombreuses techniques classiques, il est difficile pour les analystes des données, surtout ceux ne possédant pas une formation statistique poussée, de tenir compte comme il convient des effets du contrôle de la divulgation sur l'estimation. Motivés par ces limitations, nous proposons une nouvelle approche pour produire des échantillons de microdonnées à grande diffusion à partir d'enquêtes à grande échelle à laquelle nous donnons le nom de sous-échantillonnage avec synthèse. L'idée fondamentale consiste à remplacer dans l'échantillon original les valeurs identificatoires ou sensibles par des valeurs obtenues par tirages multiples à partir de modèles statistiques estimés en se servant du fichier de données originales, puis de diffuser des sous-échantillons de ces données protégées contre la divulgation. Les sous-échantillons peuvent être composés d'un ensemble commun d'enregistrements ou être tirés de manière indépendante. Cette approche est une variante de celle des données étiquetées aux États-Unis pour créer plusieurs produits de données à grande diffusion, y compris la Survey of Income

¹ Jörg Drechsler, Institut for Employment Research, Department for Statistical Methods, Regensburg Science, Box 90251, Duke University, Durham, NC 27708-0251, Allemagne. Courriel : joerg.drechsler@iab.de ; Jérôme P. Reiter, Department of Statistical Sciences, Box 90251, Duke University, Durham, NC 27708-0251. Courriel : jerry@stat.duke.edu.

- Lin, L. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, 48, 599-604.
- Lin, L., et Chinchilli, V. (1997). Rejoinder to the letter to the editor from Adkinson and Nevill. *Biometrics*, 53, 777-778.
- Liottol, N., Radaelli, T., Orsili, A., Taricco, E., Roggerol, P., Giam, M.L., Consomni, D., Moscal, F., et Cetin, I. (2010). Relationship between in utero sonographic evaluation and subcutaneous plicometry after birth in infants with intrauterine growth restriction: An exploratory study. *Italian Journal of Pediatrics*, 36, 70-77.
- Lipsitz, S.R., Laird, N.M., et Brennan, T.A. (1994). Simple moment estimates of the κ -coefficient and its variance. *Applied Statistics*, 43, 309-323.
- Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R Hoboken*. New Jersey : John Wiley & Sons, Inc.
- MacDougall, H.G., Weber, K.P., McGarvie, L.A., Halmagyi, G.M., et Curthoys, I.S. (2009). The video head impulse test. Diagnostic accuracy in peripheral vestibulopathy. *Neurology*, 73, 1134-1141.
- MacLure, M., et Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126, 161-169.
- Mazaheri, V., Hricak, H., Fine, S.W., Akin, O., Shukla-Dave, A., Ishii, N.M., Moskowitz, C.S., Grater, J.E., Reuter, V.E., Zakian, K.L., Touijer, K.A., et Koutcher, J.A. (2009). Prostate tumor volume measurement with combined T2-weighted imaging and diffusion-weighted MR: Correlation with pathologic tumor volume. *Radiology*, 252 (2), 449-457.
- National Center for Health Statistics (2011). Third National Health and Nutrition Examination Survey, 1988-1994, NHANES III Examination data file (CD-ROM). <http://www.cdc.gov/nchs/nhanes.htm>.
- National Health Interview Survey (NHIS) (2011). <http://www.cdc.gov/nchs/nhis.htm>.
- Zar, J.H. (1996). *Biostatistical Analysis*. Upper Saddle River, New Jersey : Prentice Hall International.
- Robison, W. (1999). On weighted kappa and concordance correlation coefficient. Thèse de doctorat, University of Illinois in Chicago/Graduate College/Mathematics.
- Rust, K.F., et Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Sherry, B., Jeffers, M.E., et Grummer-Strawn, L.M. (2007). Accuracy of adolescent self-report of height and weight in assessing overweight status: A literature review. *Archive of Pediatrics Adolescent Medicine*, 161, 1154-1161.
- Shield, M., Gorber, S.C., et Tremblay, M.S. (2008). Effects of measurement on obesity and morbidity. *Health Reports*, 19, 77-84.
- Slade, G.D., et Beck, J.D. (1999). Plausibility of periodontal disease estimates from NHANES III. *Journal of Public Health Dentistry*, 59, 67-72.
- Talamayan, K.S., Springer, A.E., Kelder, S.H., Gorospe, E.C., et Joye, K.A. (2006). Prevalence of overweight misperception and weight control behaviors among normal weight adolescents in the United States. *The Scientific World Journal*, 6, 365-373.
- Tanner, M.A., et Young, M.A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, 80, 175-180.
- Williamson, J.M., Manatunga, A.K., et Lipsitz, S.R. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics*, 1, 191-202.
- Wim, D.M., Johnson, C.L., et Kingman, A. (1999). Periodontal disease estimates in NHANES III: Clinical measurement and complex sample design issues. *Journal of Public Health Dentistry*, 59, 73-78.
- Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

- Étapes du calcul de kappa et de son erreur-type en utilisant un logiciel standard d'analyse de données d'enquête
- Étape 1 : Estimer les probabilités des variables binaires Y_{hi1} et Y_{hi2} en utilisant un logiciel d'analyse de données d'enquête qui intègre la stratification, la mise en grappes et la pondération de l'échantillon (par exemple PROC SURVEY-FREQ en SAS).
- Étape 2 : Estimer $P_e = (\hat{\pi}_1\hat{\pi}_2 + (1 - \hat{\pi}_1)(1 - \hat{\pi}_2))$.
- Étape 3 : Créer la nouvelle variable d'accord $U_{hi} (= Y_{hi1}Y_{hi2} + (1 - Y_{hi1})(1 - Y_{hi2}))$.
- Étape 4 : Calculer la somme des poids de sondage et la somme pondérée des U_{hi} (par exemple, en utilisant PROC SURVEYMEANS en SAS). Estimer kappa en utilisant l'équation (2).
- Étape 5 : Créer une nouvelle variable z_{hi} en utilisant l'équation (6).
- Étape 6 : Calculer l'erreur-type de z_{hi} en utilisant un logiciel standard d'analyse des données d'enquête. L'erreur-type de z_{hi} estime l'erreur-type de $\hat{\kappa}$.
- Étape 1 : Estimer les probabilités des variables binaires Y_{hi1} et Y_{hi2} en utilisant un logiciel d'analyse de données d'enquête
- Étape 2 : Estimer $P_e = (\hat{\pi}_1\hat{\pi}_2 + (1 - \hat{\pi}_1)(1 - \hat{\pi}_2))$.
- Étape 3 : Créer la nouvelle variable d'accord $U_{hi} (= Y_{hi1}Y_{hi2} + (1 - Y_{hi1})(1 - Y_{hi2}))$.
- Étape 4 : Calculer la somme des poids de sondage et la somme pondérée des U_{hi} (par exemple, en utilisant PROC SURVEYMEANS en SAS). Estimer kappa en utilisant l'équation (2).
- Étape 5 : Créer une nouvelle variable z_{hi} en utilisant l'équation (6).
- Étape 6 : Calculer l'erreur-type de z_{hi} en utilisant un logiciel standard d'analyse des données d'enquête. L'erreur-type de z_{hi} estime l'erreur-type de $\hat{\kappa}$.

Bibliographie

- Barnhart, H.X., et Williamson, J.M. (2001). Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics*, 57, 931-940.
- Behavioral Risk Factor Surveillance System (BRFSS). <http://www.cdc.gov/BRFSS>.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 20, 37-46.
- Binder, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases : une approche de type « recette ». *Techniques d'enquête*, 22, 17-22.
- Carrasco, J.L., et Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, 59, 849-858.
- Chinchilli, V.M., Mantel, J.K., Kummamkya, S. et Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, 52, 341-353.
- Cochran, W.G. (1963). *Sampling Techniques*, 2^e Ed. New York : John Wiley & Sons, Inc.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.
- Statistique Canada, N° 12-001-X au catalogue
- Dixon, A.E., Sugar, E.A., Zinreich, S.J., Slavin, R.G., Corren, J., Nacion, R.M., Ishii, M., Cohen, R.L., Brown, E.D., Wisc, R.A. et et Irvin, C.G. (2009). Criteria to screen for chronic sinusitis disease. *Chest*, 136 (5), 1324-1332.
- Femleib, M., Garttison, R.J., Fabisz, R.R., Christian, J.C., Hrubec, Z., Bottham, N.O., Kannel, W.B., Roseman, R., Schwartz, J.T. et Wagner, J.O. (1977). The NHLBI Twin Study of cardiovascular disease risk factors: Methodology and summary of results. *American Journal of Epidemiology*, 106, 284-295.
- Fieid, A.E., Anjea, P. et Rosner, B. (2007). The validity of self-reported weight change among adolescents and young adults. *Obesity*, 15, 2357-2364.
- Feis, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2^e Edition. New York : John Wiley & Sons, Inc.
- Feis, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York : John Wiley & Sons, Inc.
- Feis, J.L., Nec, J.C.M. et Landis, J.R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86, 974-977.
- Gillmour, E., Ellerbrock, T.V., Koulos, J.P., Chasson, M.A., Williamson, J.M., Kuhn, L. et Wright, T.C. (1997). Measuring cervical ectopy: Direct visual assessment versus computerized planimetry. *American Journal of Obstetrics and Gynecology*, 176, 108-111.
- Gorber, S.C., Tremblay, M., Moher, D. et Gorber, B. (2007). A comparison of direct vs. self-report measures for assessing height, weight and body mass index: A systematic review. *Obesity Review*, 8, 373-374.
- Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York : John Wiley & Sons, Inc. Vols I and II.
- Khawaja, I.S., Olson, E.J., van der Wal, C., Bukatky, J., Somers, V., Dieckhising, R. et Morgensthaler, T.L. (2010). Diagnostic accuracy of split-night polysomnograms. *Journal of Clinical Sleep Medicine*, 6 (4), 357-362.
- Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.
- Klar, N., Lipsitz, S.R. et Ibrahim, J.G. (2000). An estimating equations approach for modeling kappa. *Biometrical Journal*, 42, 45-58.
- Kocks, J.W., Kerstjens, H.A., Snijders, S.L., de Vos, B., Bierman, J.J., van Hengel, P., Strijbos, J.H., Bosveld, H.E. et van der Moelen, T. (2010). Health status in routine clinical practice: validity of the clinical COPD questionnaire at the individual patient level. Health and Quality of Life Outcomes, 8, 135-141.
- Korten, A.E., Jorm, A.F., Henderson, A.S., McCusker, E. et Creasey, H. (1992). Control-informant agreement on exposure history in case-control studies of Alzheimer's disease. *International Journal of Epidemiology*, 21, 1121-1131.
- Liang, K.Y., et Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255-268.

Étape 4 : Calculer le CCC en substituant les moyennes et les variances estimées dans l'équation (1). Créer la nouvelle variable Z_{hij} fondée sur l'équation (4).

Étape 5 : Calculer l'erreur-type de Z_{hij} en utilisant un logiciel standard d'analyse des données d'enquête. L'erreur-type de Z_{hij} estime l'erreur-type de $\hat{\rho}_c$.

CODE SAS :

```

PROC SURVEYMEANS DATA=dataset MEAN;
/*Étape 1 susmentionnée*/;
STRATA s;
CLUSTER c;
WEIGHT w;
VAR y1 y2;
ODS OUTPUT STATISTICS=stat;
data _null;
set stat (where=(varname='y1'));
call symputx('muy1', mean);
data _null;
set stat (where=(varname='y2'));
call symputx('muy2', mean);
data dataset; set dataset;
cy1 = y1 - &muy1;
cy2 = y2 - &muy2;
vary1 = cy1 ** 2;
vary2 = cy2 ** 2;
covy12 = cy1 * cy2;
PROC SURVEYMEANS MEAN;
/*Étape 3 susmentionnée*/;
STRATA s;
CLUSTER c;
WEIGHT w;
VAR vary1 vary2 covy12;
ODS OUTPUT STATISTICS=stat;
data _null;
set stat (where=(varname='vary1'));
call symputx('vary1', mean);
data _null;
set stat (where=(varname='vary2'));
call symputx('vary2', mean);
data dataset; set dataset;
d = &vary1 + &vary2 + (&muy1 - &muy2) ** 2;
CCC = 2 * &covy12 / d;
z = (2 / d) * (cy1 * cy2) - (2 * &covy12 / d) * ((cy1 ** 2) + (cy2 ** 2) + 2 * (&muy1 - &muy2) * (y1 - y2));
PROC SURVEYMEANS MEAN;
/*Étape 5 susmentionnée*/;
STRATA s;
CLUSTER c;
WEIGHT w;
VAR CCC z;
run;
```

entre les erreurs-types des estimations pondérées et non pondérées des erreurs-types, pour le CCC ainsi que le coefficient kappa. Les intervalles de confiance qui tiennent compte des poids de sondage et des caractéristiques du plan permettent d'obtenir des inférences correctes.

À l'annexe, nous montrons les étapes du calcul des mesures pondérées du CCC et du coefficient kappa ainsi que de leurs erreurs-types en utilisant un logiciel standard d'analyse de données d'enquête qui prend en compte les poids de sondage, la mise en grappes et la stratification. L'approche GEE est avantageuse parce qu'il s'agit d'un cadre commode pour élaborer des estimations des coefficients d'accord et qu'il est facile de l'étendre à des évaluations multiples, des méthodes multiples, un ajustement des covariables et des tailles de grappes non équilibrées. Cette approche fondée sur le plan de sondage donne une estimation correcte de l'erreur-type sans que l'on doive émettre des hypothèses concernant un modèle sous-jacent et tenir compte du plan d'échantillonnage. Si l'on souhaite estimer l'accord entre deux variables ordinaires à l'aide du coefficient kappa, l'approche des équations d'estimation généralisées de Williamson et coll. (2000) peut être étendue de la même façon à la méthode proposée.

Remerciements

Nous remercions les examinateurs et l'éditeur pour leurs commentaires utiles. Entre autres, l'éditeur a fourni des suggestions très judicieuses pour la section d'introduction.

Annexe

Étapes du calcul du CCC et de son erreur-type en utilisant un logiciel standard d'analyse de données d'enquête

Étape 1 : Calculer les moyennes des variables continues Y_{hij1} et Y_{hij2} en utilisant un logiciel d'analyse de données d'enquête qui intègre la stratification, la mise en grappes et la pondération de l'échantillon (par exemple PROC SURVEYMEANS en SAS).

Étape 2 : Elever au carré les valeurs de Y_{hij1} et Y_{hij2} centrées autour de leur moyenne respective.

Étape 3 : Calculer les moyennes des carrés des valeurs centrées de Y_{hij1} et Y_{hij2} en utilisant un logiciel standard pour l'analyse des données d'enquête. Ces moyennes sont les estimations de la variance de Y_{hij1} et Y_{hij2} . Calculer la moyenne du produit des valeurs centrées de Y_{hij1} et Y_{hij2} en utilisant un logiciel standard pour l'analyse de données d'enquête. Cette moyenne est la covariance d'enquête. Cette moyenne est la covariance d'enquête.

variance estimée de Y_{hij1} et Y_{hij2} .

run;

Le coefficient kappa est un choix naturel pour évaluer l'accord entre deux évaluations binaires car il comporte une correction pour l'accord dû au hasard (Fleiss 1981). L'estimation de kappa pour mesurer l'accord entre les deux définitions de la maladie parodontale (profondeur de la poche parodontale ≥ 4 mm et perte d'attachement ≥ 3 mm) en utilisant la méthode proposée est égale à 0,307. L'erreur-type de 0,0158 a été obtenue par la méthode de linéarisation de Taylor ainsi que par la méthode du jack-knife. Au tableau 2, ces résultats sont comparés aux mesures obtenues quand on ne tient pas compte du plan de sondage complexe. L'erreur-type du coefficient kappa est plus grande lorsque l'on prend en considération le plan de sondage.

Tableau 2
Moyenne non pondérée et pondérée, kappa et erreurs-types pour la perte d'attachement et la profondeur de la poche parodontale

	Perte	Profondeur	Kappa
d'attachement de la poche			
Estimation non pondérée	0,93	0,283	0,334
E.-T.	0,004	0,004	0,008
Estimation pondérée	0,358	0,212	0,307
E.-T.	0,009	0,016	0,0158

4. Discussion

Le CCC et le coefficient kappa évaluent l'accord entre deux mesures pour des réponses continues et catégorielles, respectivement. Dans le présent article, nous avons proposé une approche fondée sur des équations d'estimation généralisées pour estimer le CCC pour une paire de variables continues et le coefficient kappa pour une paire de variables binaires, d'après des données d'enquête, quand les données ont été recueillies selon un plan de sondage complexe comportant des caractéristiques telles que la stratification ou la mise en grappes. L'estimateur de type sandwich habituel de la variance ne tient compte que des mesures répétées faites sur la même personne et ne prend pas en considération le cadre d'échantillonnage (par exemple grappes, stratification et pondération). Dans l'approche GEE, les erreurs-types des estimateurs sont estimées par les méthodes de linéarisation de Taylor et du jackknife. Si les données ne sont pas recueillies selon un plan de sondage complexe, les estimateurs proposés sont identiques aux estimateurs habituels. Comme le montre les deux exemples tirés de l'étude NHANES III, il est nécessaire d'intégrer dans l'analyse les poids de sondage et les caractéristiques du plan de sondage afin que les erreurs-types ne soient pas sous-estimées quand les données sont recueillies selon un plan de sondage complexe. Les tableaux 1 et 2 révèlent des écarts importants respectivement.

coefficient de corrélation de Pearson est égal à 0,85. Dans cette situation, le CCC reflète la reproductibilité ainsi que les différences entre les moyennes des poids auto-déclarés et mesurés. Par conséquent, le CCC est informatif et avantageux si l'on s'intéresse à ces comparaisons, particulièrement dans l'analyse par domaine dans le cas d'une enquête complexe.

3.2 Étude de la santé buccodentaire

Slade et Beck (1999) ont utilisé la mesure de la profondeur de la poche parodontale et de la perte d'attachement comme indice de maladies parodontale. La prévalence de la maladie parodontale en utilisant les seuls publiés antérieurement de profondeur de la poche parodontale ≥ 4 mm et de perte d'attache ≥ 3 mm a été estimée par Slade et Beck (1999, tableau 1). La profondeur de la poche parodontale peut être le reflet d'une inflammation plutôt que d'une maladie parodontale chronique, de sorte que le niveau d'attachement de la gencive pourrait être une mesure plus significative de la destruction parodontale. Néanmoins, la profondeur de la poche parodontale demeure la mesure recommandée en pratique clinique (Wim, Johnson et Kingman 1999). Par conséquent, nous comparons l'accord de ces deux définitions de la maladie parodontale en utilisant le coefficient kappa.

Nous nous servons de l'échantillon qui a été analysé par Slade et Beck (1999). Les données portent sur 14 415 personnes de 13 ans et plus pour lesquelles une évaluation de la profondeur des poches parodontales et de la perte d'attachement avait été effectuée par six dentistes désignés. De nouveau, nous utilisons les données recueillies pour la période complète de six ans de l'enquête (1988 à 1991 ainsi que 1991 à 1994). Le nombre total de strates était de 49 et chaque strate contenait deux UPE. Nous avons utilisé dans nos analyses la variable étiquetée poids d'échantillonnage, wptx6, qui tient compte des probabilités de sélection différentes. La première définition de la maladie parodontale correspond à une profondeur de la poche parodontale ≥ 4 mm et à la deuxième, à une perte d'attachement maximale ≥ 3 mm. Pour les deux variables, nous utilisons les valeurs maximales pour l'ensemble des dents dans la bouche d'une personne. Les estimations de la probabilité pour les variables de perte d'attachement et de profondeur de la poche parodontale sont égales à 0,358 (e.-t. jackknife = 0,0088) et 0,212 (e.-t. jackknife = 0,016), respectivement, en utilisant la méthode proposée. Les erreurs-types asymptotiques fondées sur le développement en série de Taylor usuel (Woodruff 1971, produites par PROC SURVEYFREQ en SAS, version 9.1) sont égales à 0,0088 et 0,015, respectivement.

par le National Center for Health Statistics des Centers for Disease Control and Prevention, a été conçue comme une enquête de six années divisée en deux phases (1988 à 1991 et 1991 à 1994). Les données ont été recueillies suivant un plan de sondage probabiliste complexe, à plusieurs degrés, en vue de sélectionner des participants représentatifs de la population américaine civile, ne vivant pas en établissement. Des renseignements détaillés sur le plan de sondage ainsi que les lignes directrices concernant l'analyse et la diffusion des données ont été publiés dans les manuels de référence et les rapports de la NHANES III (National Center for Health Statistics 1996).

3.1 Étude du poids chez les adolescents

Le CCC est un choix naturel pour évaluer l'accord entre deux mesures du poids, parce que celles-ci sont faites sur la même échelle et que leurs éléments sont comparables (poids autodéclaré : 78 lb ~ 350 lb et poids réel : 73 lb ~ 372 lb) (Lin et Chinchilli 1997). L'estimation du CCC pour évaluer l'accord entre les deux définitions du poids en utilisant la méthode proposée est égale à 0,93. L'erreur-type de l'estimation est égale à 0,021 en utilisant la méthode de linéarisation de Taylor. L'erreur-type par jackknife de 0,021 concorde étroitement avec l'erreur-type par linéarisation. Ces statistiques sont résumées au tableau 1 ainsi que leur valeur calculée quand il n'est pas tenu compte de la structure d'échantillonnage. Les erreurs-types des estimations tenant compte du plan de sondage sont beaucoup plus grandes que les estimations non pondérées.

Tableau 1
Moyenne non pondérée et pondérée, CCC, et erreurs-types respectives pour le poids autodéclaré et le poids réel des adolescents, en livres

Autodéclaré	Réel	CCC
135,31	136,96	0,890
E.-T.	0,76	0,0005
Estimation non pondérée		
135,47	136,30	0,926
E.-T.	1,82	0,0205
Estimation pondérée		

Comme le CCC, le coefficient de corrélation de Pearson usuel entre les mesures autodéclarées et le poids réel est également égal à 0,93. Dans ce cas, l'écart moyen entre les deux mesures du poids est à peine inférieur à une livre. Lorsque l'on examine les sous-populations, on constate des différences entre le CCC et le coefficient de corrélation de Pearson. Considérons la sous-population de personnes dont le poids mesuré était supérieur à 200 lb au moment de l'examen physique. Si l'on résume les données pour cette sous-population, le poids autodéclaré est, en moyenne, inférieur de 8 livres au poids mesuré (223,2 lb c. 231,4 lb). Il existe un léger écart du CCC (0,72) par rapport au coefficient de corrélation de Pearson (0,76). La différence entre les deux mesures augmente chez le sous-groupe des personnes plus obèses. Dans la sous-population pour laquelle le poids mesuré est supérieur à 220 lb, les moyennes des poids autodéclarés et mesurés sont 231,9 lb et 248,8 lb, respectivement. Le CCC est égal à 0,67, tandis que le

coefficient de corrélation de Pearson est de 0,72. La différence entre le poids autodéclaré et le poids mesuré (en livres) chez les adolescents (de 12 à 16 ans).

L'autodéclaration du poids a été obtenue juste avant sa mesure réelle. Nous utilisons les données couvrant la période complète de six ans de l'enquête (de 1988 à 1991 et de 1991 à 1994). Pour simplifier, nous avons exclu une strate qui ne contenait qu'une seule UPE. Le nombre de strates était donc de 48 et chaque strate contenait 2 UPE. Nous avons utilisé dans les analyses le poids de sondage étiqueté wptex6 qui tient compte des probabilités de

entre le poids autodéclaré et le poids mesuré (en livres) chez les adolescents (de 12 à 16 ans).

NHANES III pour estimer le CCC qui mesure l'accord entre le poids autodéclaré et le poids mesuré (en livres) chez les adolescents (de 12 à 16 ans).

2008). Nous utilisons des données provenant de la composante des mesures corporelles (anthropométrie) de la NHANES III pour estimer le CCC qui mesure l'accord entre le poids autodéclaré et le poids mesuré (en livres) chez les adolescents (de 12 à 16 ans).

et les problèmes de santé peuvent être surestimés si l'on se sert du poids autodéclaré (Shield, Gorber et Tremblay 2008). Nous utilisons des données provenant de la composante des mesures corporelles (anthropométrie) de la NHANES III pour estimer le CCC qui mesure l'accord entre le poids autodéclaré et le poids mesuré (en livres) chez les adolescents (de 12 à 16 ans).

adulte a également révélé que les associations entre l'obésité et les problèmes de santé peuvent être surestimées si l'on se sert du poids autodéclaré (Shield, Gorber et Tremblay 2008). Nous utilisons des données provenant de la composante des mesures corporelles (anthropométrie) de la NHANES III pour estimer le CCC qui mesure l'accord entre le poids autodéclaré et le poids mesuré (en livres) chez les adolescents (de 12 à 16 ans).

entendre que les programmes de prévention de l'obésité devraient aborder la question des perceptions fausses du poids et les effets indésirables des méthodes de contrôle du poids nuisibles pour la santé, même chez les adolescents ayant un poids normal (Talamayan, Springer, Kelder, Gorospe et Joye 2006). Une étude canadienne comparable fondée sur des données de l'Enquête sur la santé dans les collectivités canadiennes de 2005 portait sur la population adulte a également révélé que les associations entre l'obésité et les problèmes de santé peuvent être surestimées si l'on se sert du poids autodéclaré (Shield, Gorber et Tremblay 2008). Nous utilisons des données provenant de la composante des mesures corporelles (anthropométrie) de la NHANES III pour estimer le CCC qui mesure l'accord entre le poids autodéclaré et le poids mesuré (en livres) chez les adolescents (de 12 à 16 ans).

normal se percevaient à tort comme ayant un excès de poids et adopteraient des comportements de contrôle du poids néfastes pour la santé (Field, Aneja et Rosner 2007 ; Gorber, Tremblay, Moher et Gorber 2007 ; Sherry, Jefferts et Grummer-Stuwn 2007). Donc, les chercheurs ont laissé entendre que les programmes de prévention de l'obésité devraient aborder la question des perceptions fausses du poids et les effets indésirables des méthodes de contrôle du poids nuisibles pour la santé, même chez les adolescents ayant un poids normal (Talamayan, Springer, Kelder, Gorospe et Joye 2006). Une étude canadienne comparable fondée sur des données de l'Enquête sur la santé dans les collectivités canadiennes de 2005 portait sur la population adulte a également révélé que les associations entre l'obésité et les problèmes de santé peuvent être surestimées si l'on se sert du poids autodéclaré (Shield, Gorber et Tremblay 2008). Nous utilisons des données provenant de la composante des mesures corporelles (anthropométrie) de la NHANES III pour estimer le CCC qui mesure l'accord entre le poids autodéclaré et le poids mesuré (en livres) chez les adolescents (de 12 à 16 ans).

$$\begin{aligned}\omega_{(h)l}^{(h)} &= w_{hl} \\ \omega_{(h)l}^{(h)} &= w_{hl} n_h / (n_h - 1) \quad \text{si } l \neq l \\ &= \omega_{(h)l}^{(h)} \quad \text{(même strate mais grappes différentes)} \\ \omega_{(h)l}^{(h)} &= 0 \quad \text{(pour la grappe supprimée).}\end{aligned}$$

L'estimateur de variance jackknife résultant pour \hat{p}^c est

$$v_J(\hat{p}^c) = \sum_{h=1}^H \left(\frac{n_h}{n} - 1 \right) \sum_{l=1}^{l-1} (\hat{p}^{c(h)l} - \hat{p}^c)^2$$

où $\hat{p}^{c(h)}$ est estimé de la même façon que \hat{p}^c , mais en utilisant les poids recalculés $\omega_{(h)}$ au lieu des poids ω . Les

estimateurs jackknife des moyennes sont calculés de la même façon.

2.4 Le coefficient kappa

Supposons qu'une paire de réponses binaires, X_{hjl}^{hjl} et Y_{hjl}^{hjl} , est observée pour le j^e membre de la l^e grappe de la h^e strate et que leurs valeurs prévues sont les probabilités π_1^{hjl} et π_2^{hjl} . De nouveau, supposons que nous estimons

la paire de réponses binaires intrasujet, Lipsitz et coll. (1994) ont montré comment se servir de deux ensembles d'équations d'estimation généralisées pour élaborer de simples estimations non itératives du coefficient κ qui peuvent être utilisées pour des données non équilibrées, parce que les estimations antérieures de kappa et de sa variance n'étaient proposées que pour des données équilibrées. Ils ont défini la variable aléatoire binaire $U_{hjl} = X_{hjl}^{hjl} + (1 - X_{hjl}^{hjl})(1 - Y_{hjl}^{hjl}) = 1$ si les deux réponses formant la paire concordent et 0 autrement. Par conséquent, $E[U_{hjl}] = P_o$, qui désigne la probabilité d'observer un accord et que l'on suppose ici être constante sur toutes les strates, grappes d'observations. Posons maintenant que $E[Y^{hjl} X^{hjl}] = P[Y^{hjl} = 1] = \omega$. La probabilité d'observer un accord peut être exprimée par $P_o = 1 - \pi_1 - \pi_2 + 2\omega$. La probabilité d'un accord dû au hasard est définie comme étant $P_e = \pi_1 \pi_2 + (1 - \pi_1)(1 - \pi_2)$ et est estimée par $\hat{P}_e = \hat{\pi}_1 \hat{\pi}_2 + (1 - \hat{\pi}_1)(1 - \hat{\pi}_2)$, où $\hat{\pi}_1$ et $\hat{\pi}_2$ sont calculées dans le premier ensemble d'équations d'estimation.

Nous pouvons calculer des estimations de κ d'après des données d'enquête en suivant l'approche décrite pour le CCC à la section 2.1. Nous pouvons intégrer les matrices de poids de sondage dans les deux ensembles d'équations d'estimation généralisées de Lipsitz et coll. (1994) pour estimer kappa. Puis, en choisissant les matrices de covariance de travail sous « indépendance » pour les deux ensembles d'équations comme dans l'approche de Lipsitz et coll. (1994), nous arrivons à l'estimation non itérative suivante de kappa pour des données d'enquête :

$$\hat{\kappa} = \frac{\sum_H \sum_{n_h} \sum_{m_{hl}} w_{hjl} \hat{P}_e - \hat{P}_e \sum_H \sum_{n_h} \sum_{m_{hl}} w_{hjl}}{\sum_H \sum_{n_h} \sum_{m_{hl}} w_{hjl} \sum_{j=1}^{j-1} U_{hjl} - \sum_H \sum_{n_h} \sum_{m_{hl}} w_{hjl} \sum_{j=1}^{j-1} \hat{P}_e} \quad (5)$$

Cet estimateur est identique à celui de Lumley (2010), qui peut être calculé en utilisant le logiciel R pour données d'enquête et la fonction `svykappa`.

On peut estimer l'erreur-type de $\hat{\kappa}$ de la même façon que celle de \hat{p}^c en utilisant la méthode de linéarisation de Taylor. Les dérivées premières de kappa par rapport à P_o, π_1 , et π_2 sont :

$$\begin{aligned}\frac{\partial \hat{P}_o}{\partial \kappa} &= \frac{1 - \hat{P}_e}{1}, \\ \frac{\partial \pi_1}{\partial \kappa} &= \frac{(1 - \hat{P}_e)^2}{(1 - \hat{P}_o)(1 - 2\pi_1)}, \\ \frac{\partial \pi_2}{\partial \kappa} &= \frac{(1 - \hat{P}_e)^2}{(1 - \hat{P}_o)(1 - 2\pi_1)}.\end{aligned}$$

Donc

$$\begin{aligned}\hat{\kappa} - \kappa &\approx \left(\frac{\partial \hat{P}_o}{\partial \kappa} \right) (\hat{P}_o - P_o) + \left(\frac{\partial \pi_1}{\partial \kappa} \right) (\hat{\pi}_1 - \pi_1) + \left(\frac{\partial \pi_2}{\partial \kappa} \right) (\hat{\pi}_2 - \pi_2) \\ &= \sum_H \sum_{n_h} \sum_{m_{hl}} w_{hjl}^* z_{hjl}, \quad \text{et} \\ \text{où } w_{hjl}^* &= w_{hjl} / \left(\sum_H \sum_{n_h} \sum_{m_{hl}} w_{hjl} \sum_{j=1}^{j-1} w_{hjl} \right)\end{aligned}$$

$$z_{hjl} = \left(\frac{\partial \hat{P}_o}{\partial \kappa} \right) U_{hjl} + \left(\frac{\partial \pi_1}{\partial \kappa} \right) X_{hjl} + \left(\frac{\partial \pi_2}{\partial \kappa} \right) Y_{hjl} - \frac{1 - \hat{P}_e}{U_{hjl} (1 - \hat{P}_o) [X_{hjl} (1 - 2\pi_2) + Y_{hjl} (1 - 2\pi_1)]} (1 - \hat{P}_e)^2.$$

En remplaçant dans (6) les paramètres par leurs estimations respectives, on traite ensuite z_{hjl} comme une variable aléatoire et on estime sa variance en utilisant un logiciel standard d'analyse de données d'enquête qui tient compte du plan de sondage. La variance de ce nouvel estimateur \hat{z}_{hjl} est une approximation de la variance de $\hat{\kappa}$. La méthode du jackknife peut également être utilisée pour estimer la variance de $\hat{\kappa}$.

3. Enquête NHANES III

Nous avons utilisé des données provenant de la troisième National Health and Nutrition Examination Survey pour illustrer notre méthode. La NHANES III, qui a été réalisée

$$\hat{\sigma}_{12}^2 = \frac{\left(\sum_H \sum_{h=1}^N \sum_{j=1}^{n_H} W_{hij12} X_{hij1} X_{hij2} \right)}{\left(\sum_H \sum_{h=1}^N \sum_{j=1}^{n_H} W_{hij12} \right)} - \hat{\mu}_1 \hat{\mu}_2.$$

2.2 Estimateur de la variance par linéarisation

Les estimateurs robustes habituels de la variance des moyennes et du CCC selon l'approche GEE ne sont pas valides ici puisqu'ils ne tiennent pas compte de la structure d'échantillonnage, ne prenant seulement en considération que la corrélation des observations faites sur le même individu. Nous proposons d'estimer l'erreur-type par la méthode de linéarisation de Taylor (Binder 1983 ; Binder 1996). Les dérivées premières de p_c (équation 1) par rapport à μ_1 , μ_2 , σ_1^2 , σ_2^2 , et σ_{12} sont :

$$\begin{aligned} \frac{\partial p_c}{\partial \mu_1} &= -\frac{D^2}{4\sigma_{12}(\mu_1 - \mu_2)}, & \frac{\partial p_c}{\partial \mu_2} &= \frac{D^2}{-4\sigma_{12}(\mu_2 - \mu_1)}, \\ \frac{\partial p_c}{\partial \sigma_1^2} &= -\frac{D^2}{2\sigma_{12}}, & \frac{\partial p_c}{\partial \sigma_2^2} &= \frac{D^2}{-2\sigma_{12}}, \\ \frac{\partial p_c}{\partial \sigma_{12}} &= \frac{D}{2}, \end{aligned}$$

où $D = \sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2$. Donc,

$$\begin{aligned} \hat{p}_c - p_c &\approx \left(\frac{\partial p_c}{\partial \mu_1} \right) (\hat{\mu}_1 - \mu_1) + \left(\frac{\partial p_c}{\partial \mu_2} \right) (\hat{\mu}_2 - \mu_2) \\ &+ \left(\frac{\partial p_c}{\partial \sigma_1^2} \right) (\hat{\sigma}_1^2 - \sigma_1^2) + \left(\frac{\partial p_c}{\partial \sigma_2^2} \right) (\hat{\sigma}_2^2 - \sigma_2^2) + \left(\frac{\partial p_c}{\partial \sigma_{12}} \right) (\hat{\sigma}_{12} - \sigma_{12}) \\ &= \frac{-4\sigma_{12}^2(\mu_1 - \mu_2)}{D^2} (\hat{\mu}_1 - \mu_1) + \frac{D^2}{-4\sigma_{12}(\mu_2 - \mu_1)} (\hat{\mu}_2 - \mu_2) \\ &+ \frac{D^2}{-2\sigma_{12}^2} (\hat{\sigma}_1^2 - \sigma_1^2) + \frac{D^2}{2\sigma_{12}^2} (\hat{\sigma}_2^2 - \sigma_2^2) + \frac{D}{2} (\hat{\sigma}_{12} - \sigma_{12}). \end{aligned}$$

L'équation ci-dessus peut être réécrite en deux parties, l'une comprenant les estimations des paramètres $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, et $\hat{\sigma}_{12}$, et l'autre, uniquement les paramètres qui ne contribuent pas à l'estimation de la variance de \hat{p}_c . Donc, la première partie devient

$$\begin{aligned} &= -\frac{D^2}{4\sigma_{12}^2(\mu_1 - \mu_2)} \hat{\mu}_1 - \frac{D^2}{4\sigma_{12}^2(\mu_2 - \mu_1)} \hat{\mu}_2 \\ &- \frac{D^2}{2\sigma_{12}^2} \hat{\sigma}_1^2 - \frac{D^2}{2\sigma_{12}^2} \hat{\sigma}_2^2 + \frac{D}{2} \hat{\sigma}_{12} \\ &= -\frac{D^2}{2\sigma_{12}^2} (2(\mu_1 - \mu_2)(\hat{\mu}_1 - \mu_1) + \hat{\sigma}_1^2 + \hat{\sigma}_2^2) + \frac{D}{2} \hat{\sigma}_{12} \\ &= -\frac{D^2}{2\sigma_{12}^2} \left(\sum_H \sum_{h=1}^N \sum_{j=1}^{n_H} 2(\mu_1 - \mu_2)(w_{hij}^* X_{hij1} - w_{hij}^* X_{hij2}) \right. \\ &\quad \left. + w_{hij}^* (X_{hij1} - \mu_1)^2 + w_{hij}^* (X_{hij2} - \mu_2)^2 \right) \\ &+ \frac{D}{2} \sum_H \sum_{h=1}^N \sum_{j=1}^{n_H} w_{hij}^* (X_{hij1} - \mu_1)(X_{hij2} - \mu_2) \end{aligned} \quad (3)$$

où $w_{hij}^* = w_{hij} / (\sum_H \sum_{h=1}^N \sum_{j=1}^{n_H} w_{hij})$. Après mise en évidence du terme de sommation, l'équation (3) devient une fonction linéaire des données qui peut alors être exprimée sous la forme $\sum_{h=1}^H \sum_{j=1}^{n_H} w_{hij}^* z_{hij}$, où

$$z_{hij} = -\frac{D^2}{2\sigma_{12}^2} (2(\mu_1 - \mu_2)(X_{hij1} - X_{hij2}) + (X_{hij1}^2 - \mu_1^2) + (X_{hij2}^2 - \mu_2^2)) + \frac{D}{2} (X_{hij1} - \mu_1)(X_{hij2} - \mu_2). \quad (4)$$

On crée alors à partir de l'équation (4) une variable aléatoire z_{hij} qui remplace les paramètres par leurs estimations respectives. La variance de ce nouvel estimateur \hat{z}_{hij} est une approximation de la variance de \hat{p}_c , qui peut être estimée en utilisant un logiciel standard d'analyse de données d'enquête (voir l'annexe).

2.3 Estimateur jackknife de la variance

Nous utilisons aussi la technique du jackknife pour estimer l'erreur-type des paramètres selon la méthode de Rust et Rao (1996, section 2.1) en vue de comparer les résultats aux estimations par linéarisation. La technique du jackknife est mise en œuvre en calculant un ensemble d'estimations répétées et en estimant la variance en se servant de ces observations provenant de la grappe en question. Les poids chaque grappe en supprimant de l'échantillon toutes les répétées. Un jeu de données répétées est créé pour les estimations. Les autres observations qui se trouvent dans la strate contenant la grappe sont augmentées d'un facteur $n_h/(n_h - 1)$. Les poids dans les autres strates restent inchangés. Donc, les nouveaux poids pour le jeu de données répétées créés en supprimant la grappe i de la strate h sont :

marginaux, même en cas de spécification incorrecte de la structure de corrélation en utilisant un estimateur « sandwich » robuste de la variance. Nous utilisons l'ap- proche GEE pour analyser les données d'enquête en inté- grant additionnellement une matrice de poids de sondage comme suit :

$$\sum_{h_i}^H \sum_{i=1}^{I-1} \mathbf{D}_i^{h_i} \mathbf{W}^{h_i} \mathbf{V}^{-1}(\mathbf{Y}^{h_i} - \pi^{h_i}(\hat{\alpha})) = \mathbf{0},$$

où \mathbf{D}^{h_i} est la matrice des dérivées partielles $d[\pi^{h_i}]/d\mu$ de dimensions $(b \times m^{h_i})$, \mathbf{W}^{h_i} est une matrice diagonale prin- cipale de dimensions $(m^{h_i} \times m^{h_i})$ constituée des poids de sondage au niveau de la personne $\mathbf{W}^{h_i} \mathbf{V}^{h_i}$ est une matrice de variance-covariance de travail de dimensions $(m^{h_i} \times m^{h_i})$ pour les réponses intégrappes, \mathbf{Y}^{h_i} est un vecteur de réponses de dimension $(m^{h_i} \times 1)$ comprenant les réponses $X^{h_{ij}}$ et $\pi^{h_i} = E[\mathbf{Y}^{h_i}]$ est éventuellement une fonction du vecteur de paramètres β de dimension $(q \times 1)$. Les équa- tions d'estimation généralisées (GEE) peuvent alors être résolues non itérativement, ce qui donne l'estimation habituelle

$$\hat{\alpha} = \left(\sum_{h_i}^H \sum_{i=1}^{I-1} \sum_{j=1}^J \mathbf{W}^{h_{ij}} X^{h_{ij}} \right) / \left(\sum_{h_i}^H \sum_{i=1}^{I-1} \sum_{j=1}^J \mathbf{W}^{h_{ij}} \right)$$

si nous estimons une moyenne commune $\pi = \beta$ ($q = 1$) et que nous utilisons une matrice de covariance de travail sous

indépendance.

Supposons qu'une paire de réponses continues, $X^{h_{ij1}}$ et $X^{h_{ij2}}$, est observée pour le j^{e} membre de la i^{e} grappe de la

niveau, supposons que nous estimons les moyennes communes μ_1 et μ_2 sans covariables pour la paire de réponses continues intrasujet, qui peuvent être estimées en utilisant l'équation d'estimation généralisée susmentionnée.

Barnhart et Williamson (2001) ont montré comment trois ensembles d'équations d'estimation généralisées peuvent être utilisés pour modéliser le CCC défini en (1) en utilisant des données corrélées. Nous étendons le deuxième en- semble de GEE de Barnhart et Williamson (2001) afin d'estimer les variances des réponses continues en intégrant

de nouveau une matrice de pondération comme suit :

$$\mathbf{V}_2(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2) = \sum_{h_i}^H \sum_{i=1}^{I-1} \mathbf{F}_i^{h_i} \mathbf{W}^{h_i} \mathbf{H}^{h_i} (\mathbf{Y}^{h_i2} - \delta_2^{h_i}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2)) = \mathbf{0},$$

où $\mathbf{F}_i^{h_i}$ est la matrice des dérivées partielles $d[\delta_2^{h_i}]/d\sigma^2$ de dimensions $(2 \times 2m^{h_i})$ avec $\sigma^2 = [\sigma_1^2, \sigma_2^2]$, \mathbf{W}^{h_i} est une matrice diagonale principale des poids d'échantillonnage au

où

$$\hat{\rho}_c = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + (\hat{\mu}_1 - \hat{\mu}_2)^2}{2\hat{\sigma}_1^2}$$

tivement :

où $\mathbf{C}_i^{h_i}$ est un vecteur de dérivées partielles $\partial \theta^{h_i}/\partial \rho_c$ de dimension $(1 \times m^{h_i})$, \mathbf{W}^{h_i} est une matrice diagonale prin- cipale de dimensions $(m^{h_i} \times m^{h_i})$ comprenant les poids d'échantillonnage au niveau de la personne $\mathbf{W}^{h_{ij}}$, et \mathbf{K}^{h_i} est une matrice de covariance de travail de dimensions $(m^{h_i} \times m^{h_i})$ que nous choisissons comme matrice de covariance sous « indépendance ». Les équations d'estimation généra- lisées susmentionnées peuvent être résolues non itéra-

pour la p^{e} mesure dans la paire, $p = 1, 2$. Le CCC peut être estimé au moyen d'un troisième ensemble d'équations d'estimation en utilisant les produits par paire des réponses pour modéliser σ_{12} , une fois que les moyennes et les variances sont estimées. Soit $\mathbf{U}^{h_i} = [X^{h_{i12}}, X^{h_{i21}}, X^{h_{i22}}, \dots, X^{h_{im^{h_i}2}}]$ un vecteur de dimension $(m^{h_i} \times 1)$ des produits de paires des réponses et écrivons $\theta^{h_i} = E[\mathbf{U}^{h_i}]$, qui est une fonction des moyennes, des variances et du CCC. Pour trouver la valeur de $\hat{\rho}_c$, nous résolvons un troisième ensemble d'équations d'estimation :

$$\mathbf{V}_3(\hat{\rho}_c, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2) = \sum_{h_i}^H \sum_{i=1}^{I-1} \mathbf{C}_i^{h_i} \mathbf{W}^{h_i} \mathbf{K}^{h_i} (\mathbf{U}^{h_i} - \theta^{h_i}(\hat{\rho}_c, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2)) = \mathbf{0},$$

$$\hat{\sigma}_2^p = \left(\sum_{h_i}^H \sum_{i=1}^{I-1} \sum_{j=1}^J \mathbf{W}^{h_{ij}} X_2^{h_{ij}} \right) / \left(\sum_{h_i}^H \sum_{i=1}^{I-1} \sum_{j=1}^J \mathbf{W}^{h_{ij}} \right) - \hat{\mu}_2^p$$

alors être résolues non itérativement :

niveau de la personne \mathbf{W}^{h_i} de dimensions $(2m^{h_i} \times 2m^{h_i})$, \mathbf{H}^{h_i} est une matrice de variance-covariance de travail de dimensions $(2m^{h_i} \times 2m^{h_i})$ pour les cartes des réponses intégrappes, $\mathbf{Y}_2^{h_i} = [X_2^{h_{i12}}, X_2^{h_{i21}}, X_2^{h_{i22}}, \dots, X_2^{h_{im^{h_i}2}}]$ un vecteur de réponses continues de dimen- sions $(2m^{h_i} \times 1)$, et $\delta_2^{h_i} = E[\mathbf{Y}_2^{h_i}]$. Bien que $\delta_2^{h_i}$ soit une fonction des termes de variance σ_1^2 et σ_2^2 ainsi que des moyennes μ_1 et μ_2 , nous nous supposons que les moyennes sont fixes dans $\delta_2^{h_i}$ et nous ne calculons les dérivées de $\delta_2^{h_i}$ que par rapport aux variances. De nouveau, nous choisissons la matrice \mathbf{H}^{h_i} de dimensions $(2m^{h_i} \times 2m^{h_i})$ comme matrice de variance-covariance de travail sous « indépendance », ainsi que le vecteur colonne $\delta_2^{h_i} = [\sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \dots, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2]$ de dimension $(2m^{h_i} \times 1)$ parce que nous supposons que les variances et les moyennes sont communes dans toutes les strates et toutes les grappes. Les équations d'estimation généralisées susmentionnées peuvent alors être résolues non itérativement :

de sondage dans les deux exemples. Nous concluons par une brève discussion.

2. Méthodes

Nous proposons une approche générale d'estimation du CCC et du coefficient kappa d'après des données d'enquête en suivant deux méthodes GEE. Pour le CCC, trois ensembles d'équations d'estimation sont nécessaires. Un premier ensemble modélise la distribution des réponses cont- nues. Comme l'ont fait Barnhart et Williamson (2001), nous utilisons un deuxième ensemble d'équations d'estimation pour estimer les variances des réponses continues. Enfin, le troisième ensemble sert à estimer le CCC en modélisant la covariance entre les réponses continues apparées et les estimations des moyennes d'après les deux premiers ensembles d'équations d'estimation. Pour κ , deux ensembles seulement d'équations d'estimation sont nécessaires. Un premier ensemble modélise la distribution marginale des réponses binaires. Comme dans Lipsitz, Laird et Brennan (1994), un deuxième ensemble d'équations d'esti- mation est introduit pour estimer κ en modélisant une variable aléatoire binaire qui dépeint l'accord entre deux réponses sur un sujet.

Afin de tenir compte des probabilités de sélection vari- ables, des matrices de pondération sont incorporées dans chaque ensemble d'équations d'estimation. L'estimation de l'erreur-type des estimations \hat{p}_c et $\hat{\kappa}$ proposées d'après les données d'enquête est effectuée par la méthode de linéari- sation de Taylor. Nous montrons aussi comment l'estima- tion des erreurs-types des estimateurs proposés peut se faire par l'approche du jackknife.

Supposons qu'une enquête est réalisée selon un plan de sondage avec stratification, mise en grappes et probabilités de sélection inégales. Soit Y_{hij} la variable réponse pour le j^{e} membre ($j = 1, \dots, m_h$) de la i^{e} grappe ($i = 1, \dots, n_h$) de la h^{e} strate ($h = 1, \dots, H$). En calculant la moyenne sur tous les échantillons possibles, la valeur espérée correspon- dante est $E[Y_{hij}] = \mu_{hij}$ si Y_{hij} est une réponse continue, et la probabilité correspondante est $E[Y_{hij}] = \Pr[Y_{hij} = 1] = \pi_{hij}$ si Y_{hij} est une réponse binaire. Le poids de sondage w_{hij} est l'inverse de la probabilité de sélection du j^{e} membre de la i^{e} grappe de la h^{e} strate.

2.1 Le coefficient de corrélation de concordance

Liang et Zeger (1986) ont élaboré des méthodes fondées sur les moments pour analyser les observations corrélées provenant d'une même grappe (par exemple mesures répé- tées au cours du temps sur le même individu ou observa- tions sur plusieurs membres de la même famille). L'ap- proche GEE donne une estimation cohérente des paramètres

d'obésité. Nous donnons des mesures pondérées du CCC et du coefficient kappa, ainsi que les estimateurs de variance de ces mesures en tenant compte du plan de sondage. À la section 2, nous présentons une approche fondée sur des équations d'estimation généralisées (GEE pour *generalized estimating equations*) pour estimer ces deux coefficients d'accord d'après des données d'enquête. À la section 3, nous illustrons notre méthode au moyen de données re- cueillies dans le cadre de la NHANES III. Nous utilisons des mesures corporelles pour estimer p_c en vue d'évaluer l'accord entre le poids autodéclaré et le poids réel. Nous pourrions aussi évaluer l'accord entre deux défini- tions de la maladie périodontique. Nous tenons compte de la stratification et de la mise en grappes, et intégrerons les poids pondérés des erreurs-types des estimations des paramètres dans les études par sondage. Omettre d'inclure les poids de pondération des erreurs-types des estimations des paramètres peut entraîner des estimations biaisées et non efficaces. Nous présentons les estimations pondérées et non pondérées des erreurs-types des estimations des paramètres dans les études par sondage. Omettre d'inclure les poids de sondage et de tenir compte de la structure du plan de sondage dans les analyses donne lieu à une sous-estimation des erreurs-types et à une inférence incorrecte. Cet aspect est particulièrement important dans le cas d'enquêtes répé- tées tous les deux ou trois ans, et les chercheurs s'intéressent souvent tout spécialement à la comparaison des variations entre domaines ou sous-populations. Ainsi, dans la première application de la NHANES III, nous comparons l'accord entre le poids corporel autodéclaré et le poids corporel mesuré durant l'examen physique chez les adolescents. Il est nécessaire de calculer des erreurs-types (intervalles de confiance) exactes si l'on veut comparer le CCC pour divers domaines, tels que les sous-groupes de poids normal et d'obésité.

Nous présentons les deux exemples tirés de la NHANES III présentés à la section 3, d'importants écarts peuvent exister entre les estimations pondérées et non pondérées des erreurs-types des estimations des paramètres dans les études par sondage. Omettre d'inclure les poids de sondage et de tenir compte de la structure du plan de sondage dans les analyses donne lieu à une sous-estimation des erreurs-types et à une inférence incorrecte. Cet aspect est particulièrement important dans le cas d'enquêtes répé- tées tous les deux ou trois ans, et les chercheurs s'intéressent souvent tout spécialement à la comparaison des variations entre domaines ou sous-populations. Ainsi, dans la première application de la NHANES III, nous comparons l'accord entre le poids corporel autodéclaré et le poids corporel mesuré durant l'examen physique chez les adolescents. Il est nécessaire de calculer des erreurs-types (intervalles de confiance) exactes si l'on veut comparer le CCC pour divers domaines, tels que les sous-groupes de poids normal et d'obésité.

l'indice d'apnées-hypopnées (IAH) pour les deux et trois premières heures de sommeil en considérant l'IAH pro- venant de la polysomnographie (PSG) de nuit complète comme la norme de référence. Dans le domaine de la recherche en radiologie, les associations entre les mesures du volume d'une tumeur de la prostate d'après les résultats de l'examen pathologique (Mazaheri et coll. 2009). Des tests de l'égalité du coefficient kappa ont été utilisés pour comparer l'évaluation visuelle et la planimétrie informatisée dans l'évaluation de l'ectopie cervicale (Gillmour, Ellerböck, Koulos, Chlasson, Williamson, Kuhn et Wright 1997; Williamson, Manatunga et Lipsitz 2000), et pour compa- rer les taux de cholestérol chez les jumeaux monozygotes et dizygotes (Feinleib, Gartson, Fabsitz, Christian, Hrubec, Borhani, Kannel, Roseman, Schwartz et Wagner 1977).

variance pour démontrer que le CCC coïncide avec le coefficient de corrélation intraclass (CCI) si l'on tient compte de l'écart entre les moyennes des méthodes :

$$P_{\text{ICC}} = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_e^2 + \sigma_\eta^2} = \frac{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2}.$$

Par conséquent, on peut estimer le CCC en utilisant les composantes de variance d'un modèle à effets mixtes ou la méthode habituelle des moments. En raison de sa supériorité par rapport au coefficient de corrélation de Pearson et de son lien avec le CCI, l'application du CCC a gagné en popularité ces dernières années (Chinchilli, Martel, Kumanyika et Lloyd 1996; Zar 1996). En 2009 et en 2010, le CCC a été utilisé comme mesure d'accord dans plus de 60 publications médicales dans des domaines tels que la maladie respiratoire (Dixon, Sugar, Zinreich, Slavin, Corren, Nacletio, Ishii, Cohen, Brown, Wise et Irvin 2009; Kocks, Kersjens, Snijders, de Vos, Biermann, van Hengel, Strijbos, Bosveld et van der Molén 2010), le sommeil (Khawaja, Olson, van der Wal, Bukatyk, Somers, Dieckhising et Morgensthaler 2010), la pédiatrie (Liotot, Raadell, Orsi, Tartocco, Roggero, Giam, Consonni, Mosca et Cetin 2010), la neurologie (MacDougall, Weber, McGarvie, Halmagyi et Cuthroys 2009) et la radiologie (Mazaheri, Hrtak, Fine, Akin, Shukla-Devue, Ishii, Moskowitz, Grater, Reuter, Zakian, Touijer et Kouchner 2009).

Le coefficient kappa (κ) (Cohen 1960) et le coefficient kappa pondéré (Cohen 1968) sont les indices les plus utilisés pour mesurer l'accord entre des résultats de type discret ou ordinal, respectivement (Fleiss 1981). Soient Y_1 et Y_2 deux variables aléatoires binaires prenant les valeurs 0 et 1 avec les probabilités $\pi_1 = \text{Pr}(Y_1 = 1)$ et $\pi_2 = \text{Pr}(Y_2 = 1)$. Kappa corrige le pourcentage d'accord des évaluations en tenant compte de la proportion de l'accord due au hasard (calculée sous hypothèse d'indépendance) et est défini comme suit :

$$\kappa = \frac{P_o - P_e}{1,0 - P_e}, \quad (2)$$

où P_o est la probabilité que les réponses binaires formant la paire soient égales sous hypothèse d'indépendance ($\pi_1\pi_2 + (1 - \pi_1)(1 - \pi_2)$) et P_e est la probabilité que les réponses binaires formant la paire soient égales (Cohen 1960). La différence $P_o - P_e$ représente l'excès d'accord par rapport à l'accord dû au hasard. Une valeur de 0 pour κ indique qu'il n'y a pas d'accord au-delà de celui dû au hasard et une valeur de 1 indique un accord parfait (Fleiss 1981). Les inconvénients du kappa tiennent au fait qu'il s'agit d'une fonction de la distribution marginale des évaluateurs (Fleiss, Nee et Landis 1979; Tanner et Young 1985) et que son

La valeur des enquêtes par sondage est généralement Henderson, McCusker et Creasey 1992). L'informant dans les études cas-témoins (Korten, Jorm, giques (Machure et Willett 1987) et de l'accord témoin- (Klar, Lipsitz et Ibrahim 2000), de divers outils épidémiolo- validité et la reproductibilité de la similitude entre jumeaux scores entiers. Le kappa a été employé pour mesurer la est équivalent au kappa pondéré lorsque l'on utilise des culé d'après des données mesurées sur une échelle ordinale et coll. 1979). Robieson (1999) a remarqué que le CCC cal- éendue dépend du nombre d'évaluations par sujet (Fleiss

fonde sur le plan de sondage concernant la population cible dans le cas de plans de sondage complexes, les estimateurs et leurs variances comprennent les poids de sondage et tiennent compte de la structure du plan en vue d'obtenir des estimations sans biais. En outre, grâce à l'intégration des poids de sondage et de la structure du plan de sondage dans les analyses, toute correction possible provenant des grappes dans un plan à plusieurs degrés est prise en compte, de sorte que les erreurs-types des estimateurs ne sont pas sous-estimées.

En général, les chercheurs n'essaient pas de vérifier si leur estimation de l'accord entre les mesures obtenue au moyen du CCC ou du coefficient kappa diffère de manière significative de zéro. Ils cherchent à publier des intervalles de confiance avec leurs estimations (par exemple Dixon et coll. 2009; Mazaheri et coll. 2009). Comme dans le cas du coefficient de corrélation de Pearson, il n'existe aucune valeur cible permettant de juger de la force de l'accord. Par conséquent, il est essentiel de porter le jugement quant à l'accord entre toute méthode d'essai et toute méthode de référence avec un degré de certitude établi. Dans certaines situations, les études qui sont réalisées requièrent la vérification d'hypothèses ou la comparaison d'indices d'accord pour plus d'une nouvelle méthode par rapport à une méthode de référence. Ainsi, Khawaja et coll. (2010) ont vérifié l'égalité de deux CCC calculés pour comparer

Estimation de coefficients d'accord d'après des données d'enquête

Hung-Mo Lin, Hae-Young Kim, John M. Williamson et Virginia M. Lesser¹

Résumé

Nous présentons une approche fondée sur des équations généralisées pour estimer le coefficient de corrélation de concordance et le coefficient kappa d'après des données d'enquête. Les estimations ainsi que leurs erreurs-types doivent tenir compte correctement du plan d'échantillonnage. Nous présentons des mesures pondérées du coefficient de corrélation de concordance et du coefficient kappa, ainsi que la variance de ces mesures tenant compte du plan d'échantillonnage. Nous utilisons la méthode de linéarisation par série de Taylor et la procédure du jackknife pour estimer les erreurs-types des estimations résultantes des paramètres. Des mesures anthropométriques et des données sur la santé bucco-dentaire provenant de la Third National Health and Nutrition Examination Survey sont utilisées pour illustrer cette méthodologie.

Mots clés : Mise en grappes ; coefficient de corrélation de concordance ; équations d'estimation généralisées ; estimateur jackknife ; coefficient kappa ; pondération de l'échantillon ; stratification ; linéarisation par série de Taylor.

1. Introduction

Les enquêtes comprennent souvent la collecte de mesures multiples d'états latents, tels que la qualité de vie ou le désir de poursuivre des études collégiales, ainsi que de mesures multiples d'états difficiles à classer, tels que le syndrome de fatigue chronique. Quand plusieurs mesures sont recueillies, il est naturel de vouloir évaluer leur accord et obtenir des intervalles de confiance pour ces mesures de l'accord. Il pourrait aussi être intéressant de comparer le degré d'accord chez divers sous-groupes de la population ou pour diverses paires de mesures. Dans ce contexte, on pourrait souhaiter tester l'égalité des mesures d'accord. Le présent article porte sur deux mesures de l'accord entre des mesures multiples de ce genre, le coefficient de corrélation de concordance (CCC, ρ_c) et le coefficient kappa (κ). Le premier est utile pour des mesures continues sur des échelles naturelles. Si la mesure d'un concept latent ne possède pas d'échelle naturelle, elle peut être recalculée arbitrairement de manière que sa moyenne soit nulle et sa variance égale à l'unité. Quand cela est possible, il est dénué de sens de parler d'écart entre les moments marginaux. Cependant, s'il existe une échelle naturelle, le recalcul n'est pas souhaitable et une bonne mesure de l'accord tiendra compte à la fois de la corrélation et de l'accord des moments marginaux. Le coefficient kappa est des plus utiles pour les classifications binaires.

Le CCC, comme on l'a montré, convient mieux que le coefficient de corrélation de Pearson (ρ) pour mesurer l'accord ou la reproductibilité (Lin 1989 ; Lin 1992). Il évalue l'exactitude entre deux lectures en mesurant la

variation de la relation linéaire ajustée par rapport à la droite à 45° passant par l'origine (la droite de concordance), ainsi que la précision en mesurant la quelle distance chaque observation s'écarte de la droite ajustée. Soit X_{i1} et X_{i2} une paire de variables aléatoires continues mesurées sur le même sujet i en utilisant deux méthodes. Le CCC pour mesurer l'accord de X_{i1} et X_{i2} est défini comme suit :

$$\rho_c = 1 - \frac{E[(X_{i1} - X_{i2})^2]}{2\sigma_{i2}^2} = \frac{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}{2[\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2]} \quad (1)$$

où $\sigma_1^2 = \text{var}(X_{i1})$, $\sigma_2^2 = \text{var}(X_{i2})$ et $\sigma_{i2}^2 = \text{cov}(X_{i1}, X_{i2})$ (Lin 1989). Comme l'a souligné Lin (1989), $\rho_c = 0$ si et uniquement si $\rho = 0$. On peut également montrer algébriquement que ρ_c est proportionnel à ρ et que $-1 \leq -|\rho| \leq \rho_c \leq |\rho| \leq 1$ (Lin 1989). D'où l'imprécision peut être reflétée par une valeur plus faible de ρ et un biais systématisé peut être reflété par un ratio de ρ_c à ρ plus faible. Ensemble, les renseignements sur ρ et ρ_c constituent un jeu d'outils pour déterminer quelle correction, en vue d'améliorer l'exactitude et/ou la précision, est la plus avantageuse (Lin et Chinnai 1997).

Le coefficient de corrélation intraclass (CCI) est aussi une mesure de l'accord très répandue pour les variables mesurées sur une échelle continue (Fleiss 1986). Supposons que X_{i1} et X_{i2} peuvent être décrites dans un modèle linéaire comme suit : $y_{ij} = \mu_j + \theta_j + e_{ij}$ où μ_j est la moyenne de la mesure fournie par la j^{e} méthode, $\theta_j \sim (0, \sigma_\theta^2)$ est la variable latente pour le i^{e} sujet, et les $e_{ij} \sim (0, \sigma_e^2)$ sont des termes d'erreur indépendants. Carrasco et Jover (2003, page 850) ont utilisé un modèle comprenant des composantes de

1. Hung-Mo Lin, Department of Anesthesiology, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1010, New York, NY 10029, États-Unis. Courriel : hung-mo.lin@mountsinai.org; Hae-Young Kim, Center for Statistical Analysis and Research, New England Research Institute, 9 Galen Street, Watertown, MA 02472, États-Unis; John M. Williamson, Center for Global Health Research, Centers for Disease Control and Prevention/Kenya Medical Research Institute, 1578 Kisumu-Busia Road, Kisumu, Kenya; Virginia M. Lesser, Department of Statistics and Survey Research Center, Oregon State University, 44 Kiddier Hall, Corvallis, OR 97731-4606, États-Unis.

- Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland. Available at <http://drum.lib.umd.edu/bitstream/1903/7598/1/umimd-4863.pdf>.
- Li, J., et Valliant, R. (2009). Matrice chapéau et effets de levier pondérés par les poids de sondage. *Techniques d'enquête*, 35(1), 17-27.
- Li, J., et Valliant, R. (2011). Linear regression influence diagnostics for unclustered survey data. *Journal of Official Statistics*, 20, 99-119.
- Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Thèse de doctorat non-publiée, University of Maryland. Disponible au http://drum.lib.umd.edu/bitstream/1903/10881/1/Liao_umd_0117E_11537.pdf.
- Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.
- Neter, J., Kutner, M., Wasserman, W. et Nachtsheim, C. (1996). *Applied Linear Statistical Models*. New York : McGraw-Hill/Irwin, 4^e Ed.
- Simon, S.D., et Lesage, J.P. (1988). The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management*, 1, 137-152.
- Smith, G. (1974). Multicollinearity and forecasting. Yale University Cowles Foundation Discussion Paper No.383. Disponible au <http://cowles.econ.yale.edu/P/cd/d03b/d0383.pdf>.
- Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.
- Theil, H. (1971). *Principles of Econometrics*. New York : John Wiley & Sons, Inc.

$$a^{(k)}_k = \mathbf{I}^T_k \mathbf{A}^{-1} \mathbf{I}_k = \mathbf{I}^T_k (\mathbf{X}^T_k \mathbf{X}^{(k)}_k)^{-1} \mathbf{I}_k$$
$$= \frac{(1 - R^2_{\text{ppS}(k)})}{1 - R^2_{\text{ppS}(k)} (\text{SST}^{\text{ppS}(k)})} = \frac{(1 - R^2_{\text{ppS}(k)})}{1 - R^2_{\text{ppS}(k)} (\mathbf{X}^T_k \mathbf{X}^{(k)}_k)}$$
$$R^2_{\text{ppS}(k)} = \frac{\mathbf{\hat{b}}^{\text{ppS}(k)}_T \mathbf{X}^T_k \mathbf{X}^{(k)}_k \mathbf{\hat{b}}^{\text{ppS}(k)}_T}{\text{SST}^{\text{ppS}(k)}},$$

où

avec $\mathbf{\hat{b}}^{\text{ppS}(k)}_T = (\mathbf{X}^T_k \mathbf{X}^{(k)}_k)^{-1} \mathbf{X}^T_k \mathbf{y}^{(k)}_k$ est le coefficient de détermination correspondant à la régression de $\mathbf{y}^{(k)}_k$ sur les $p - 1$ autres variables explicatives. Le terme $\text{SST}^{\text{ppS}(k)} = \mathbf{y}^{(k)T}_k \mathbf{y}^{(k)}_k$, est la somme totale des carrés des écarts dans cette régression.

Le terme $(1 - R^2_{\text{ppS}(k)})^{-1}$ dans (16) est le VIF qui sera produit par les projectiels standard quand est exécutée une régression par les moindres carrés pondérés. Sous le modèle $\mathbf{y} = \mathbf{X}\mathbf{\beta} + \epsilon$ avec $\epsilon \sim (0, \sigma^2 \mathbf{W}^{-1})$, l'expression (16) est égale à $\text{Var}(\mathbf{\hat{b}}^{\text{ppS}(k)}) / \sigma^2$. Cependant, elle n'est pas appropriée pour les régressions par les moindres carrés pondérés par les poids de sondage, parce que la variance de $\mathbf{\hat{b}}^{\text{ppS}}$ possède la forme plus complexe donnée par (2).

La matrice $\mathbf{G} = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$ peut être exprimée sous la forme :

$$\mathbf{G} = \begin{pmatrix} a^{(k)}_{kk} & \mathbf{a}^{(k)}_{(k)} \\ b^{(k)}_{kk} & \mathbf{b}^{(k)}_{(k)} \end{pmatrix} \begin{pmatrix} a^{(k)}_{kk} & \mathbf{a}^{(k)}_{(k)} \\ \mathbf{b}^{(k)}_{(k)} & \mathbf{B}^{(k)(k)} \end{pmatrix} \begin{pmatrix} a^{(k)}_{kk} & \mathbf{a}^{(k)}_{(k)} \\ \mathbf{b}^{(k)}_{(k)} & \mathbf{B}^{(k)(k)} \end{pmatrix} \quad (17)$$

où la matrice inverse $\mathbf{A}^{-1} = [a^{(k)}_{hk}]$, $h, k = 1, \dots, p$, $\mathbf{a}^{(k)}_{(k)}$ est définie comme étant la k^e ligne de \mathbf{A}^{-1} en excluant $a^{(k)}_{kk}$, $(a^{(k)}_{k1}, \dots, a^{(k)}_{k(k-1)}, a^{(k)}_{kp}, \dots, \mathbf{a}^{(k)}_{(k)(k)})^T$ et $\mathbf{A}^{(k)(k)}$ est définie comme étant la partie $(k - 1) \times (k - 1)$ de la matrice \mathbf{A}^{-1} en excluant les k^e ligne et colonne. La version partitionnée de \mathbf{B} est

$$\mathbf{B} = \begin{pmatrix} b^{(k)}_{kk} & \mathbf{b}^{(k)}_{(k)} \\ b^{(k)}_{(k)k} & \mathbf{B}^{(k)(k)} \end{pmatrix} = \begin{pmatrix} \mathbf{z}^T_k \mathbf{V} \mathbf{z}_k & \mathbf{z}^T_k \mathbf{V} \mathbf{X}^{(k)}_k \\ \mathbf{X}^{(k)T}_T \mathbf{V} \mathbf{z}_k & \mathbf{X}^{(k)T}_T \mathbf{V} \mathbf{X}^{(k)}_k \end{pmatrix} \quad (18)$$

En vertu de la symétrie de \mathbf{A} et de \mathbf{B} , le k^e élément diagonal de \mathbf{G} est

$$g^{(k)}_{kk} = a^{(k)}_{kk} (a^{(k)}_{kk} b^{(k)}_{kk} + 2\mathbf{b}^{(k)}_{(k)k} \mathbf{a}^{(k)}_{(k)k}) + \mathbf{a}^{(k)}_{(k)kT} \mathbf{B}^{(k)(k)} \mathbf{a}^{(k)}_{(k)k} \quad (19)$$

En utilisant l'inverse partitionnée de la matrice \mathbf{A} , qui représente $(\mathbf{X}^T_k \mathbf{X}^{(k)}_k)^{-1}$, on peut montrer que

$$\mathbf{a}^{(k)}_{(k)} = -a^{(k)}_{kk} (\mathbf{X}^T_k \mathbf{X}^{(k)}_k)^{-1} \mathbf{X}^{(k)}_T \mathbf{y}^{(k)}_k = -a^{(k)}_{kk} \mathbf{\hat{b}}^{\text{ppS}(k)}_T \quad (20)$$

En substituant $a^{(k)}_{(k)}$ dans (19), $g^{(k)}_{kk}$ peut être exprimé de manière compacte en fonction de $a^{(k)}_{kk}$, $\mathbf{\hat{b}}^{\text{ppS}(k)}_T$ et la composante inférieure droite de la matrice \mathbf{B} :

$$g^{(k)}_{kk} = (a^{(k)}_{kk})^2 (b^{(k)}_{kk} - 2\mathbf{b}^{(k)}_{(k)kT} \mathbf{\hat{b}}^{\text{ppS}(k)}_T + \mathbf{\hat{b}}^{\text{ppS}(k)}_T \mathbf{B}^{(k)(k)} \mathbf{\hat{b}}^{\text{ppS}(k)}_T) \\ + a^{(k)}_{kk} \times \frac{1}{1 - R^2_{\text{ppS}(k)}} \times \frac{\mathbf{X}^T_k \mathbf{y}^{(k)}_k}{1 - R^2_{\text{ppS}(k)}} \\ \times (\mathbf{z}^T_k \mathbf{V} \mathbf{z}_k - 2\mathbf{z}^T_k \mathbf{V} \mathbf{X}^{(k)}_k \mathbf{\hat{b}}^{\text{ppS}(k)}_T + \mathbf{\hat{b}}^{\text{ppS}(k)}_T \mathbf{V} \mathbf{X}^{(k)}_k \mathbf{\hat{b}}^{\text{ppS}(k)}_T)$$

$$\text{où } \hat{\mathbf{e}}^{(k)}_k = \mathbf{y}^{(k)}_k - \mathbf{X}^{(k)}_k \mathbf{\hat{b}}^{\text{ppS}(k)}_T \text{ est le résidu de la régression de } \mathbf{y}^{(k)}_k \text{ sur } \mathbf{X}^{(k)}_k \quad (21)$$

Bibliographie

Belsey, D.A. (1984). Collinearity and forecasting. *Journal of Forecasting*, 38, 73-93.

Belsey, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York : John Wiley & Sons, Inc.

Belsey, D.A., Kuh, E. et Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York : Wiley Interscience.

Elliot, M. (2007). Réduction bayésienne des poits pour les modèles de régression linéaire généralisée. *Techniques d'enquête*, 33, 27-40.

Farrar, D.E., et Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.

Fox, J. (1984). *Linear Statistical Models and Related Methods, With Applications to Social Research*. New York : John Wiley & Sons, Inc.

Fox, J., et Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417), 178-183.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28(1), 5-25.

Kmenta, J. (1986). *Elements of Econometrics*. New York : Macmillan, 2^e Ed.

Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3341-3348.

Tableau 3
Valeurs du VIF pour trois méthodes de régression différentes

Variable	MCO	MCP	VIF	Modèle complet	
				VIF	MCP
Âge	1,02	1,03	1,07	0,96	0,94
Race noire	1,10	1,07	1,12	1,05	1,05
Calories	3 411,61	3 562,70	2 740,83	0,77	0,81
Protéines	123,12	127,35	103,50	0,46	0,69
Glucides	1 074,87	1 007,40	462,08	0,78	1,70
Sucres	8,37	7,03	4,87	0,60	0,60
Fibres	4,59	3,94	2,37	0,60	0,60
Alcool	120,56	115,67	89,92	0,78	1,70
Lipides totaux	1 190,24	1 475,27	2 513,69	1,80	1,80
Acides gras saturés totaux	76,80	112,61	202,91	2,67	2,39
Acides gras mono-insaturés totaux	82,37	107,34	286,24	2,67	2,39
Acides gras poly-insaturés totaux	34,73	49,45	118,21	2,39	2,39

Variable	MCO	MCP	VIF	Modèle réduit	
				VIF	MCP
Âge	1,00	1,00	0,98	0,98	0,98
Race noire	1,02	1,01	0,97	0,96	0,96
Lipides totaux	20,10	20,22	63,15	3,12	3,04
Acides gras mono-insaturés totaux	20,16	20,26	61,57	3,04	3,04

Variable	MCO	MCP	VIF	Modèle réduit	
				VIF	MCP
Âge	1,00	1,00	0,98	0,97	0,97
Race noire	1,00	1,03	1,00	1,00	1,00
Calories	1,00	1,01	0,96	0,95	0,95

Annexe A

Détermination de g_{kk}

De façon analogue à la détermination du VIF pour les MCO classiques dans Theil (1971), la somme des carrés des écarts et la matrice des produits croisés $A = X^T X$, qui peut

$$A_{p \times p} = \begin{pmatrix} X_T^T X_K & X_T^T X_{(k)} \\ X_K^T X_T & X_K^T X_{(k)} \end{pmatrix} \quad (15)$$

où les colonnes de X sont réordonnées de manière que $X = (X_K X_{(k)})$ avec $X_{(k)}$ désignant la matrice de dimensions $n \times (p - 1)$ contenant toutes les colonnes sauf la k^e colonne de X .

En utilisant la formule de l'inverse d'une matrice partitionnée, l'élément supérieur gauche de A^{-1} peut être exprimé par :

Les objectifs de l'analyse doivent être pris en considération pour décider de la façon d'utiliser les VIF. Si l'objectif principal est la prédiction, le fait d'inclure des variables collinéaires ou de choisir des variables incorrectes n'est pas très préoccupant. Si des conclusions plus fondamentales sont souhaitées, l'analyste devrait examiner quelles variables il serait logique d'inclure comme variables explicatives au lieu de se fier à un algorithme automatique pour sélectionner les variables. Les VIF sont un outil utile pour repérer les variables explicatives dont les coefficients estimés ont une variance inutilement grande. Quoiqu'ils puissent être considérés comme un outil pour la sélection automatique des variables, les simulations présentées dans Liao (2010), non reproductibles ici, montrent que leur utilisation n'est pas un moyen fiable d'identifier un vrai modèle sous-jacent.

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs dont les commentaires ont apporté des améliorations importantes à l'article.

Figure 1 Diagrammes de dispersion par paire et coefficients de corrélation des variables nutritionnelles^a



Tableau 1 : Composition chimique des aliments d'origine animale					
Variable		Béta	MCO	Béta	E.T.
Ordonnée à l'origine		63,90*** ^d	6,95	67,47***	6,36
Âge		0,26	0,19	0,08	0,18
Race noire		10,39***	2,07	10,59***	2,38
Calories		-6,41	5,76	-8,19	5,56
Protéines		25,72	24,76	40,98	23,60
Glucides		26,67	23,93	32,31	22,96
Sucres		-1,90	3,06	-0,30	2,82
Fibres		-41,17	20,23	-34,20	17,98
Lipides totaux		38,84	39,45	49,37	38,28
Acides gras saturés totaux		150,25*	69,53	161,78*	72,12
Acides gras mono-insaturés totaux		-113,20*	49,81	-101,40	56,26
Acides gras poly-insaturés totaux		-92,60*	46,13	-75,55	51,16
Modèle complet					
		Béta	MCP	Béta	E.-T.
Ordonnée à l'origine		67,52***	6,29	67,52***	6,29
Âge		0,27	0,19	0,07	0,18
Race noire		12,54***	1,98	11,74***	2,32
Calories		0,15	0,10	0,23*	0,09
Modèle réduit					
		Béta	MCP	Béta	E.-T.
Ordonnée à l'origine		62,26***	6,88	67,52***	6,29
Âge		0,27	0,19	0,07	0,18
Race noire		12,54***	1,98	11,74***	2,32
Calories		0,15	0,10	0,23*	0,09
MCPPS					
		Béta	MCPPS	Béta	E.-T.
Ordonnée à l'origine		67,52***	8,48	67,52***	8,48
Âge		0,27	0,25	0,07	0,25
Race noire		12,54***	2,05	11,74***	2,05
Calories		0,15	0,10	0,23*	0,10

5. Conclusion

Afin qu'ils soient applicables à des modèles estimés d'après des données d'enquête, les diagnostics de régression doivent être adaptés de manière à tenir compte de la pondération et des caractéristiques du plan d'échantillonnage, telles que la stratification et la mise en grappes. Dans le présent article, nous avons élaboré une nouvelle formule pour calculer un facteur d'inflation de la variance (VIF) approprié pour les modèles linéaires. Un VIF est une mesure de l'augmentation, ou inflation, de la variance de l'estimateur d'un paramètre attribuable au fait que les variables explicatives sont corrélées au lieu d'être orthogonales. Bien que les projections standard permettent de produire des estimations de la pente d'une droite de régression pondérée par les poids de sondage en suivant des procédures qui reposent sur les moindres carrés pondérés, les VIF calculés par les routines développées pour des données ne provenant pas d'enquêtes sont incorrects. Le VIF pour un échantillon complexe est égal au VIF issu des moindres carrés pondérés multiplié par un facteur d'ajustement. Ce dernier est positif, mais peut être plus grand ou plus petit que 1, selon la nature des données analysées.

Dans une étude empirique, nous avons illustré l'application de notre nouvelle approche en utilisant des données provenant de la National Health and Nutrition Examination Survey de 2007-2008. Nous avons donné un exemple simple de la façon dont la colinéarité entre les variables explicatives influence l'estimation des coefficients dans la régression linéaire et démontre que, même si les coefficients estimés (et les valeurs prédites) sont les mêmes quand on utilise les moindres carrés pondérés ou les moindres carrés pondérés par les poids de sondage, leurs variances estimées et leurs valeurs du VIF (qui reflètent l'effet de la colinéarité sur l'estimation du coefficient) peuvent être différentes.

variables explicatives sont incluses dans le modèle, la variable des calories obtient la valeur du VIF la plus élevée dans toutes les régressions en raison de sa forte quasi-dépendance avec toutes les autres variables d'apport nutritionnel total. Comme le montre le tableau 1, le VIF pour les MCPs peut être obtenu en multipliant le VIF pour les MCP par le coefficient d'ajustement $\zeta_k p_k$. Au tableau 3, les coefficients d'ajustement $\zeta_k p_k$ pour toutes les variables d'apport nutritionnel sauf les lipides sont inférieurs à 1, d'apport nutritionnel et les lipides sont inférieurs à 1, surtout celui des glucides, qui est de 0,46. Cela indique que les MCPs ne contiennent que des variables d'apport nutritionnel : les lipides totaux et les acides gras mono-insaturés totaux. Les valeurs du VIF sous les MCPs sont trois fois plus grandes que sous les MCO ou les MCP pour ces deux variables nutritionnelles. Si un analyste applique la méthode des MCPs aux données d'enquête examinées ici et qu'il utilise les valeurs non ajustées du VIF fournies par les projections statistiques standard pour les MCO (pré-sentées dans la première colonne) ou pour les MCP (pré-sentées dans la deuxième colonne), son appréciation de la gravité de la colinéarité dans le modèle sera quelque peu erronée. Brièvement, même si les pentes estimées et les pré-dictions de la régression par les MCP et par les MCPs sont les mêmes, les VIF risquent d'être sous-estimés ou surestimés si l'on ne tient pas compte des aspects complexes de l'enquête.

Tableau 1
Méthodes de régression et leurs statistiques de diagnostic de la colinéarité utilisées dans l'étude expérimentale

Type de régression	Matrice de pondération W	Estimation de la variance de β	Formule du VIF
MCO	I	$\sigma^2(X^T X)^{-1}$	$VIF = \frac{1}{1 - R^2_{m(k)}}$
MCP	W^c	$\sigma^2(X^T W X)^{-1}$	$VIF = \frac{1 - R^2_{PPS(m(k))}}{1}$
MCPs	W	$\sigma^2(X^T W X)^{-1} X^T W V W X (X^T W X)^{-1}$	$VIF = \frac{1 - R^2_{PPS(m(k))}}{1 - \zeta_k^2 p_k}$
avec			
$\hat{\zeta}_k = \frac{e_{xk}^T W V W e_{xk}}{e_{xk}^T W e_{xk}},$			
$p_k = \frac{(\hat{x}_k - 1\hat{x}_k)^T V (\hat{x}_k - 1\hat{x}_k)}{(\hat{x}_k^T \hat{x}_k - N\hat{x}_k^T \hat{x}_k)}$			
$V = \sum_{h=1}^H \frac{m_h}{m_h - 1} \left[\text{Bld}(\text{diag}(e_{mh}^T e_{mh}) - \frac{1}{m_h} e_{mh} e_{mh}^T) \right]$			
Dans tous les modèles de régression, les paramètres sont estimés par : $\hat{\beta} = (X^T W X)^{-1} X^T W Y$.			
$R^2_{m(k)}$ est le R-carré de la régression par les MCO de x_k sur les x dans la partie restante de X (en excluant une colonne pour l'ordonnée à l'origine).			
W^c est la matrice diagonale qui contient les poids de sondage w_i sur la diagonale principale.			

Center for Health Statistics des États-Unis recommande que le plan de sélection de l'échantillon s'approche de la sélection stratifiée avec remise de 32 UPB dans 16 strates, avec 2 UPB dans chaque strate.

Pour notre étude empirique, nous avons ajusté un modèle de régression linéaire du poids corporel (en kg) par la méthode des moindres carrés pondérés par les poids de sondage. Les variables explicatives considérées comprennent l'âge, la race noire et les variables d'apport nutritionnel total quotidien, qui sont les calories (100 kcal), les lipides totaux (100 g), les acides gras saturés totaux (100 g), les sucres (100 g), les glucides (100 g), les protéines (100 g), les acides gras mono-insaturés totaux (100 g), les acides gras poly-insaturés totaux (100 g) et l'alcool (100 g). Toutes les variables d'apport nutritionnel total quotidien sont corrélées les unes aux autres à divers degrés comme l'illustre la figure 1.

Trois méthodes de régression ont été comparées dans l'étude. La première s'appuie sur la méthode des *moindres carrés ordinaires* (MCO) et ne tient pas compte des aspects complexes de l'échantillonnage, y compris la pondération. La deuxième est la méthode des *moindres carrés pondérés* (MCP), dans laquelle sont intégrées les poids de sondage en supposant que $V = W^{-1}$, mais qui ne tient pas compte de tous les aspects complexes de l'échantillonnage. La troisième est la méthode des *moindres carrés pondérés par les poids de sondage* (MCPPS), qui s'appuie sur le plan d'échantillonnage complexe réel comme il est décrit à la section 3.4. Les matrices des poids, les estimateurs des variances des coefficients et les diagnostics de collinéarité de ces trois méthodes sont présentés au tableau 1.

Les résultats de l'ajustement du modèle en utilisant les trois méthodes de régression différentes sont présentés au tableau 2. Le modèle contenant toutes les variables explicatives figure dans la partie supérieure du tableau. Dans le tiers inférieur du tableau, un modèle réduit, dans lequel le problème de quasi-dépendance est moindre, est ajusté au moyen de trois variables explicatives seulement : l'âge, la race noire et les calories. Dans le modèle réduit, la valeur du coefficient pour les calories est positive et significative quand les MCP ou les MCPPS sont utilisés, ce qui paraît logique et reflète la relation positive attendue entre le poids corporel d'une répondante et son apport calorifique total quotidien. Cependant, lorsque l'on inclut les autres variables d'apport nutritionnel total dans le modèle, la valeur du coefficient des calories est négative et non significative à cause de l'inflation de sa variance. Il s'agit d'un exemple de type dans lequel il y a inflation de la variance d'un coefficient et où son signe est illogique à cause de la collinéarité.

Le tableau 3 donne les valeurs du VIF pour les trois méthodes de régression utilisées. Les formules du VIF pour ces méthodes sont présentées au tableau 1. Quand toutes les

Nous allons maintenant illustrer les diagnostics de collinéarité modifiés proposés et étudier leur comportement en utilisant des données sur l'apport alimentaire provenant de la National Health and Nutrition Examination Survey (NHANES) réalisée aux États-Unis en 2007-2008 (http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/data/doc_changes_0708.htm). Les données sur l'apport alimentaire sont utilisées pour estimer les types et les quantités d'aliments et de boissons consommés durant la période de 24 heures qui précède l'entrevue (de minuit à minuit) et pour estimer les apports d'énergie, de nutriments et d'autres composants alimentaires provenant de ces aliments et boissons. La NHANES est réalisée selon un plan d'échantillonnage probabiliste à plusieurs degrés complexe. Certains sous-groupes de population sont suréchantillonnés afin d'accroître la fiabilité et la précision des estimations des indicateurs de l'état de santé pour ces groupes. Parmi les participants qui répondent à l'interview sur place au centre d'examen mobile (CEM), environ 94 % fournissent des renseignements complets sur les apports alimentaires. Les poids de sondage pour ces données ont été construits en prenant les poids de sondage ajustés pour le CEM et en les rajustant en outre pour tenir compte de la non-réponse supplémentaire et de la différence de répartition selon le jour de la semaine pour la collecte des données sur les apports alimentaires. Ces derniers poids sont plus variables que les poids produits pour le CEM. Le jeu de données utilisé dans notre étude est un sous-ensemble des données de 2007-2008 composé de femmes de 26 à 40 ans ayant répondu à l'enquête. Les observations comportant des valeurs manquantes pour les variables choisies ont été exclues de l'échantillon qui, en dernière analyse, contient 672 réponses complètes. Les poids finaux dans notre échantillon varient de 6 028 à 330 067, avec un ratio de 55 pour 1. Le National

4. Étude expérimentale

Étant donné ces estimateurs des composantes, VIF_k est estimé par

$$\hat{p}_{mk} = \frac{(x_k^T W x_k - \bar{1} x_k^T W \bar{1} W x_k - 1 x_k^T)}{(x_k^T W x_k - N \bar{x}_k^2)} \quad (14)$$

avec $e_{x_k} = x_k - X^{(k)} \hat{\beta}^{PPS(k)}$. L'estimation de \hat{p}_{mk} , défini selon (8), est

$$\hat{p}_k = \frac{x_k^T W x_k}{x_k^T W \bar{1} W x_k},$$

$$\widehat{VIF}_k = \frac{1 - R_k^2}{\hat{p}_k} \quad \text{et } \widehat{VIF}_{mk} \text{ est estimé par}$$
$$\widehat{VIF}_{mk} = \frac{1 - R_{mk}^2}{\hat{p}_{mk}}.$$

$$\frac{\hat{V}_{\min}}{\hat{V}} = \frac{w_{\max}}{w_{\min}}.$$

L'intervalle de $\zeta_k p_{mk}$ dépend aussi de l'intervalle des poids de sondage, comme celui de $\zeta_k p_k$.

3.4 Estimation du VIF pour un modèle avec échantillonnage en grappes stratifié quand V est inconnue

Aux sections qui précèdent, nous avons utilisé des arguments fondés sur le modèle pour calculer les VIF. Ces derniers contiennent des termes, V en particulier, qui sont inconnus et doivent être estimés. À la présente section, nous construisons des estimateurs des composantes du VIF, de nouveau en nous servant d'arguments fondés sur le modèle. Cependant, un estimateur par linéarisation classique de la variance sous le plan permet aussi d'estimer la variance sous le modèle, comme nous le montrons plus bas, et fournit les composantes nécessaires pour estimer le VIF. Dans la suite de la présente section, nous présentons des estimateurs qui conviennent pour un modèle possédant une structure de covariance en grappes stratifiée.

Supposons que, dans un plan d'échantillonnage à plusieurs degrés stratifié, il existe $h = 1, \dots, H$ strates dans la population, $i = 1, \dots, N^h$ grappes dans la strate hi . Nous sélectionnons $i = 1, \dots, n^h$ grappes dans la strate hi . L'ensemble de grappes échantillonnées dans la strate h par s_h et l'échantillon d'unités dans la grappe hi par s_{hi} . Le nombre total d'unités échantillonnées dans la strate h est $m_h = \sum_{i \in s_h} m_{hi}$, et le total dans l'échantillon est $m = \sum_{h=1}^H m_h$. Nous supposons que les grappes sont sélectionnées avec remise dans les strates et indépendamment entre les strates. Considérons le modèle qui suit :

$$\begin{aligned} E_{hi}(Y^{hi}) &= x_{hi}^T \beta \\ Y^{hi} &= 1, \dots, N^h, H, \quad i = 1, \dots, N^h, \quad h = 1, \dots, H, \\ \text{Cov}_{hi}(Y^{hi}, Y^{h'i'}) &= 0 \quad h \neq h', \quad h = h' \text{ et } i \neq i'. \end{aligned} \quad (9)$$

β pps = $\sum_{h=1}^H \sum_{i \in s_h} V^{-1} x_{hi}^T W_{hi}^T x_{hi}^T W_{hi}$ (10)
 W_{hi} est la matrice de dimensions $m_{hi} \times p$ des covariables pour les unités échantillonnées dans la grappe hi , $W_{hi} = \text{diag}(w_i)$, $i \in s_{hi}$ est la matrice diagonale des poids

de sondage pour la grappe hi et V_{hi} est le vecteur de dimensions $m_{hi} \times 1$ de variables réponses dans la grappe hi . La variance sous le modèle de β pps est :

$$\text{Var}_{hi}(\beta^{\text{pps}}) = V^{-1} \sum_{h=1}^H \sum_{i \in s_h} x_{hi}^T W_{hi}^T V_{hi} W_{hi} x_{hi} = V^{-1} \sum_{h=1}^H \sum_{i \in s_h} x_{hi}^T W_{hi}^T V_{hi} W_{hi} x_{hi} \quad (11)$$

Désignons les résidus au niveau de la grappe comme un vecteur, $e_{hi} = Y_{hi} - X_{hi} \beta^{\text{pps}}$. Un estimateur par linéarisation sous le plan de sondage est donné par :

$$\begin{aligned} \text{var}_L(\beta^{\text{pps}}) &= V^{-1} \sum_{h=1}^H \sum_{i \in s_h} \left((z_{hi} - \bar{z}_{hi})(z_{hi} - \bar{z}_{hi})^T \right) = V^{-1} \sum_{h=1}^H \sum_{i \in s_h} z_{hi} z_{hi}^T \\ &= V^{-1} \sum_{h=1}^H \sum_{i \in s_h} \left(z_{hi} z_{hi}^T - n^h \bar{z}_{hi} \bar{z}_{hi}^T \right) = V^{-1} \sum_{h=1}^H \sum_{i \in s_h} z_{hi} z_{hi}^T \quad (12) \end{aligned}$$

et $z_{hi} = X_{hi}^T W_{hi} e_{hi}$ avec $e_{hi} = Y_{hi} - X_{hi} \beta^{\text{pps}}$. Cette expression peut être réduite à la formule pour un plan de sondage stratifié à seul degré quand les tailles de échantillons de grappes sont toutes égales à 1, $m_{hi} = 1$. L'expression (12) est utilisée par les logiciels Stata et SUDAAN, entre autres. L'estimateur $\text{var}_L(\beta^{\text{pps}})$ est convergent et approximativement sans biais par rapport au plan sous un plan dans lequel les grappes sont sélectionnées avec remise (Fuller 2002). Li (2007a, b) a montré que (12) est également un estimateur approximativement sans biais par rapport au modèle sous le modèle (11).

Le terme entre crochets dans (12) sert d'estimateur de la matrice B dans l'expression (2). Les composantes de $\text{var}_L(\beta^{\text{pps}})$ peuvent être utilisées pour construire les estimateurs de ζ_k et p_k dans (5) et de p_{mk} dans (8). En particulier,

$$\begin{aligned} \zeta_k &= \frac{e_{jk}^T W_{jk} V e_{jk}}{e_{jk}^T W_{jk} V e_{jk}}, \quad \text{où} \\ V &= \text{Blikg} \left[\frac{n^h - 1}{n^h} V^h - \frac{1}{n^h} e^h e^h{}^T \right], \quad h = 1, 2, \dots, H \\ \text{avec } V^h &= \text{Blikg}(\text{diag}(e_{hi}^h e_{hi}^h{}^T)) \text{ et} \end{aligned} \quad (13)$$

du ratio des formes quadratiques (Lin 1984), les bornes du terme ζ_k sont celles de l'intervalle $\mu^{\min}(\mathbf{V}) \leq \zeta_k \leq \mu^{\max}(\mathbf{V})$, et les bornes de p_k , celles de l'intervalle

$$1 \leq p_k \leq \frac{\mu^{\max}(\mathbf{V})}{1},$$

où $\mu^{\min}(\mathbf{V})$ et $\mu^{\max}(\mathbf{V})$ sont les valeurs singulières minimale et maximale de la matrice \mathbf{V} . En combinant ces résultats, le coefficient conjoint $\zeta_k p_k$ a pour bornes l'intervalle :

$$\frac{\mu^{\min}(\mathbf{V})}{\mu^{\max}(\mathbf{V})} \leq \zeta_k p_k \leq \frac{\mu^{\max}(\mathbf{V})}{\mu^{\min}(\mathbf{V})}.$$

Notons que, quand $\mathbf{V} = \mathbf{I}$, $\zeta_k = p_k = 1$ et (6) se réduit à

$$1 - R_{\text{PPS}(k)}^2 = \frac{1}{\sigma^2} \mathbf{x}_k^T \mathbf{W} \mathbf{x}_k,$$

qui est la variance sous le modèle de l'estimation par les MCP quand \mathbf{V} est diagonale et que \mathbf{W} est correctement spécifiée comme étant $\mathbf{W} = \mathbf{V}^{-1}$. Dans ce cas inhabituel, le VIF calculé les projectifs courants sera approprié pour les MCPs. Cependant, l'hypothèse que $\mathbf{W} = \mathbf{V}^{-1}$ est rarement raisonnable dans l'estimation d'après des données d'enquête. Si $\mathbf{V} \neq \mathbf{I}$, alors ζ_k et p_k ne sont pas égaux à 1 et un calcul spécial du VIF demeure nécessaire. Quand $\mathbf{V} = \mathbf{I}$, soit l'application habituellement considérée par les analystes,

$$\mathbf{V} = \mathbf{W}, \zeta_k = \frac{\mathbf{e}_T^T \mathbf{W} \mathbf{e}_k}{\mathbf{e}_T^T \mathbf{W} \mathbf{x}_k}, p_k = \frac{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k},$$

et

$$\frac{\mu^{\min}(\mathbf{V})}{\mu^{\max}(\mathbf{V})} = \frac{w_{\min}}{w_{\max}},$$

où w_{\min} est la valeur minimale des poids de sondage et w_{\max} est leur valeur maximale. Dans ce cas, l'intervalle de $\zeta_k p_k$ est borné par

$$\left[\frac{w_{\min}}{w_{\max}}, \frac{w_{\max}}{w_{\min}} \right].$$

Quand tous les poids de sondage sont constants, $\zeta_k p_k = 1$ et les VIF produits par les logiciels standard, $(1 - R_{\text{PPS}}^2)^{-1}$, n'ont pas besoin d'être ajustés sous les MCPs; cependant, si l'intervalle des poids de sondage est grand, $\zeta_k p_k$ peut être très petit ou très grand, et peut être supérieur ou inférieur à 1. Dans ce cas, le VIF produit par les logiciels standard ne convient pas et un calcul spécial est nécessaire. Ces faits seront illustrés dans nos études expérimentales. Le VIF donné par (6) est approprié, que le modèle commun ou non une ordonnée à l'origine. Une autre version peut être écrite en supposant que le modèle contient une

Liao (2010). Nous résumons le résultat ci-dessous. La variance de $\beta_{\text{PPS}(k)}^2$ dans un modèle M2 qui contient une ordonnée à l'origine et dans lequel \mathbf{x}_k est orthogonale aux autres \mathbf{x} est :

$$\text{Var}_{M(2)}(\beta_{\text{PPS}(k)}^2) = \frac{\text{SST}_{\text{PPS}(k)}^2}{(\mathbf{x}_k - \frac{1}{N} \mathbf{1}_k)^T \mathbf{V} (\mathbf{x}_k - \frac{1}{N} \mathbf{1}_k)} \quad (7)$$

où $\mathbf{1} = (w_{1/2}, \dots, w_{n/2})^T$, $\mathbf{x}_k = \sum_{i \in S} w_i x_{ki} / N$, $N = \sum_{i \in S} w_i$, et $\text{SST}_{\text{PPS}(k)}^2 = \mathbf{x}_k^T \mathbf{x}_k - N \bar{x}_k^2$. La variance de $\beta_{\text{PPS}(k)}^2$ peut alors s'écrire

$$\text{Var}_M(\beta_{\text{PPS}(k)}^2) = \frac{1 - R_{\text{PPS}(k)}^2}{\zeta_k^2 p_{mk}^2} \text{Var}_{M(2)}(\beta_{\text{PPS}(k)}^2) \quad (8)$$

où $R_{\text{PPS}(k)}^2$ est le R-carré de la régression par les MCPs de \mathbf{x}_k sur les \mathbf{x} dans la partie restante de \mathbf{X} (en excluant une colonne pour l'ordonnée à l'origine). Le terme ζ_k a été défini en suivant (5) et

$$p_{mk} = \frac{(\mathbf{x}_k - \frac{1}{N} \mathbf{1}_k)^T \mathbf{V} (\mathbf{x}_k - \frac{1}{N} \mathbf{1}_k)}{(\mathbf{x}_k^T \mathbf{x}_k - N \bar{x}_k^2)}.$$

La plupart des projectifs donnent systématiquement $(1 - R_{\text{PPS}(k)}^2)^{-1}$ comme la sortie de la régression par les MCP. Il convient de souligner que ce paramètre est différent du VIF égal à $(1 - R_{\text{PPS}(k)}^2)^{-1}$ présenté à la section 3.3, qui ne suppose pas qu'une ordonnée à l'origine est gardée dans le modèle. Habituellement, les projectifs ne fournissent pas $(1 - R_{\text{PPS}(k)}^2)^{-1}$.

Au moyen d'arguments similaires à ceux de la section précédente, nous pouvons borner $\zeta_k p_{mk}$ par

$$\frac{\mu^{\min}(\mathbf{V})}{\mu^{\max}(\mathbf{V})} \leq \zeta_k p_{mk} \leq \frac{\mu^{\max}(\mathbf{V})}{\mu^{\min}(\mathbf{V})}.$$

La variance sous le modèle de $\beta_{\text{PPS}(k)}^2$ est augmentée d'une valeur

$$\text{VIF}_{mk} = \frac{1 - R_{\text{PPS}(k)}^2}{\zeta_k^2 p_{mk}^2}$$

comparativement à sa variance sous le modèle (M2) contenant uniquement la variable explicative \mathbf{x}_k et l'ordonnée à l'origine. Le nouveau VIF VIF_{mk} ajusté pour tenir l'ordonnée à l'origine garde certaines propriétés du VIF $_k$ donné en (6). Quand $\mathbf{V} = \mathbf{I}$, nous avons $\zeta_k = 1$, $p_{mk} = 1$ et le VIF corrigé pour l'ordonnée à l'origine dans (8) pour les MCPs est égal au VIF corrigé pour l'ordonnée à l'origine classique : $(1 - R_{\text{PPS}(k)}^2)^{-1}$. Quand $\mathbf{V} = \mathbf{I}$, nous avons $\mathbf{V} = \mathbf{W}$,

$$\zeta_k = \frac{\mathbf{e}_T^T \mathbf{W} \mathbf{e}_k}{\mathbf{e}_T^T \mathbf{W} \mathbf{x}_k}, p_{mk} = \frac{(\mathbf{x}_k - \frac{1}{N} \mathbf{1}_k)^T \mathbf{W} (\mathbf{x}_k - \frac{1}{N} \mathbf{1}_k)}{(\mathbf{x}_k^T \mathbf{x}_k - N \bar{x}_k^2)}.$$

sur un modèle, est la méthode des moindres carrés pondérés (MCP). La plupart des logiciels statistiques courants (par exemple, SAS, Stata, S-Plus et R) utilisent $(1 - R_{k(MCP)}^2)^{-1}$ comme VIF pour les MCP, où $R_{k(MCP)}^2$ est le carré de la corrélation multiple provenant de la régression par les MCP de la k^e colonne de \mathbf{X} sur les autres colonnes. Fox et Monette (1992) ont également généralisé ce concept d'inflation de la variance en tant que mesure de colinéarité à un sous-ensemble de paramètres dans \mathbf{b} et dérivé un facteur d'inflation de la variance généralisé (GVIF pour *generalized variance-inflation factor*). De surcroît, certains travaux intéressants ont abouti à l'élaboration de mesures de type VIF, telles que les *indices de colinéarité* de Stewart (1987) qui sont simplement les racines carrées des VIF et la *tolérance* définie comme étant l'inverse du VIF dans Simon et Lesage (1988).

3. Le VIF dans la régression par les moindres carrés pondérés par les poids de sondage

3.1 Estimateurs par les moindres carrés pondérés par les poids de sondage

Supposons que le modèle structurel sous-jacent à la superpopulation est $\mathbf{Y} = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{e}$, où les termes d'erreur du modèle ont une structure de variance générale $\mathbf{e} \sim (0, \sigma^2 \mathbf{V})$ avec \mathbf{V} et σ^2 connues. Soit \mathbf{W} la matrice diagonale des poids de sondage. Nous supposons tout au long de l'exposé que les poids de sondage sont construits de telle façon qu'ils peuvent être utilisés pour estimer les totaux de population finie. L'estimateur par les moindres carrés pondérés par les poids de sondage (MCPPS) est $\hat{\boldsymbol{\beta}}^{pps} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$, en supposant que $\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X}$ est inversible. Fuller (2002) décrit les propriétés de cet estimateur. L'estimateur $\hat{\boldsymbol{\beta}}^{pps}$ est un estimateur sans biais par rapport au modèle de $\boldsymbol{\beta}$ sous le modèle $\mathbf{Y} = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{e}$, que $\text{Var}_M(\mathbf{e}) = \sigma^2 \mathbf{V}$ soit spécifiée correctement ou non, et est apparemment sans biais sous le plan pour le paramètre de population (recensement) $\mathbf{B}_U = (\mathbf{X}_U^T \mathbf{X}_U)^{-1} \mathbf{X}_U^T \mathbf{Y}_U$, dans la population finie de N unités. L'indice U désigne la population finie, $\mathbf{Y}_U = (Y^1, \dots, Y^N)^T$, et $\mathbf{X}_U = (\mathbf{x}^1, \dots, \mathbf{x}^N)$ avec \mathbf{x}_k le vecteur de dimension $N \times 1$ des valeurs de la covariable k .

La variance sous le modèle de l'estimateur du paramètre $\hat{\boldsymbol{\beta}}^{pps}$, en supposant que $\text{Var}_M(\mathbf{e}) = \sigma^2 \mathbf{V}$, peut s'exprimer

$$\text{Var}_M(\hat{\boldsymbol{\beta}}^{pps}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} \sigma^2 = \mathbf{G} \sigma^2, \quad (2)$$

3.2 Variance des estimations des coefficients sous le modèle

Comme nous le montrons à l'annexe A, la variance sous le modèle de $\hat{\boldsymbol{\beta}}^{pps}$, dans (4) peut s'écrire :

$$\text{Var}_M(\hat{\boldsymbol{\beta}}^{pps}) = \mathbf{G}_{kk} \sigma^2 = \frac{1 - R_{2(pps(k))}^2}{\zeta_k^k \mathbf{x}_k^T \mathbf{V} \mathbf{x}_k} \frac{(\mathbf{x}_k^T \mathbf{x}_k)^2}{2}, \quad (5)$$

$$\zeta_k^k = \frac{\mathbf{e}_k^T \mathbf{V} \mathbf{e}_k}{\mathbf{e}_k^T \mathbf{W} \mathbf{V} \mathbf{W} \mathbf{e}_k} = \frac{\mathbf{e}_k^T \mathbf{W} \mathbf{e}_k}{\mathbf{e}_k^T \mathbf{W} \mathbf{V} \mathbf{W} \mathbf{e}_k},$$

avec $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}}^{pps(k)}$ représentant le résidu de la régression par les MCPPS de \mathbf{x}_k sur \mathbf{X}_k , et $\hat{\mathbf{e}}_{xk} = \mathbf{x}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}}^{pps(k)}$ est $\mathbf{W}_{1/2} \mathbf{e}_{xk}$.

$$\rho_k^k = \frac{\mathbf{x}_k^T \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k} = \frac{\mathbf{x}_k^T \mathbf{V} \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W} \mathbf{V} \mathbf{W} \mathbf{x}_k}$$

et $R_{2(pps(k))}^2$, défini à l'annexe A, est le carré de la corrélation multiple provenant de la régression pondérée de la k^e colonne de \mathbf{X} sur les autres colonnes. Donc, ζ_k^k et ρ_k^k dépendent de \mathbf{W} et de \mathbf{V} . La variance sous orthogonalité donnée par (3) est augmentée

$$\text{VIF}_k^k = \frac{1 - R_{2(pps(k))}^2}{\zeta_k^k \rho_k^k} \quad (6)$$

fois lorsque l'on introduit les $p - 1$ autres variables explicatives dans les MCPPS. Dans ces derniers, le VIF fondé sur le modèle inclut non seulement le coefficient de corrélation multiple $R_{2(pps(k))}^2$ mais aussi deux coefficients d'ajustement, ζ_k^k et ρ_k^k , qui ne sont pas présents dans le cas des MCO et des MCP. En procédant à la décomposition de \mathbf{V} en valeurs singulières, nous pouvons borner le facteur $\zeta_k^k \rho_k^k$, qui est l'ajustement du VIF dans les MCP. Sur la base des extrema

\mathbf{X} et une variable d'analyse Y . Le modèle linéaire classique, dans un contexte autre qu'une enquête, est $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, où \mathbf{Y} est un vecteur de dimension $n \times 1$ d'observations sur une variable réponse, ou variable dépendante; $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ est une matrice de plan de dimensions $n \times p$ de constantes fixes avec \mathbf{x}_k , le vecteur de dimension $n \times 1$ des valeurs de la variable explicative k pour les n unités échantillonnées; $\boldsymbol{\beta}$ est un vecteur de dimension $p \times 1$ des paramètres à estimer; et $\boldsymbol{\epsilon}$ est un vecteur de dimension $n \times 1$ de termes d'erreur statistiquement indépendants de moyenne nulle et de variance constante σ^2 . Nous supposons, pour simplifier, que \mathbf{X} est de plein rang-colonne. L'estimation par les moindres carrés ordinaires (MCO) de $\boldsymbol{\beta}$ est $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, pour lequel la variance sous le modèle est $\text{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Ici, nous utilisons l'indice M pour désigner l'espérance sous le modèle.

Les colinéarités des variables explicatives accroissent la variance des coefficients de régression du modèle comparativement à la situation où les \mathbf{X} sont orthogonales. On peut voir cet effet dans la formule de la variance de l'estimation d'un coefficient de régression particulier autre que l'ordonnée à l'origine β_k (Theil 1971),

$$(1) \quad \text{Var}_M(\hat{\beta}_k) = \frac{\sigma^2}{1 - R_k^2} \sum_{i \in S} \frac{x_{ik}^2}{x_{ik}^2} \quad 1 - R_k^2$$

où R_k^2 est le carré de la corrélation multiple provenant de la régression de la k^{e} colonne de \mathbf{X} sur les autres colonnes. Ce R-carré, défini comme étant $R_k^2 = \beta_{(k)}\mathbf{X}_{(k)}'\mathbf{X}_{(k)}\beta_{(k)}/\mathbf{x}_k'\mathbf{x}_k$ ou $\beta_{(k)}$ est l'estimation par les MCO de la pente quand on fait la régression de \mathbf{x}_k sur les autres \mathbf{x} et que $\mathbf{X}_{(k)}$ est la matrice \mathbf{X} dont la k^{e} colonne a été supprimée. Le terme $\sigma^2/\sum x_{ik}^2$ est la variance de $\hat{\beta}_k$ sous le modèle si la k^{e} variable explicative est orthogonale à toutes les autres variables explicatives. La valeur de R_k^2 peut être non nulle parce que la k^{e} variable explicative est corrélée à une autre variable explicative ou à cause d'une relation plus complexe de dépendance entre \mathbf{x}_k et plusieurs autres variables explicatives. Par conséquent, la colinéarité entre \mathbf{x}_k et certaines autres variables explicatives peut entraîner l'inflation de la variance de $\hat{\beta}_k$ au-delà de la valeur qui serait obtenue si les \mathbf{x} étaient orthogonales. Le deuxième terme de (1), $(1 - R_k^2)^{-1}$, est appelé facteur d'inflation de la variance, symbolisé par VIF, pour *variance inflation factor* (Theil 1971).

Une référence fondamentale sur la colinéarité et d'autres diagnostics de la régression par les MCO est le document de Belsley et coll. (1980). Les diagnostics de colinéarité sont abordés dans de nombreux autres traités, y compris Fox (1984) et Neter et coll. (1996). Dans certains cas, il est souhaitable de pondérer les cas différemment dans une analyse de régression afin d'intégrer une variance résiduelle non constante. Cette forme de pondération, qui est fondée

diagnostiquer la présence de colinéarité (par exemple, Farrar Fox 1984; Belsley 1991), aucun de ces travaux de recherche ne traite des diagnostics de colinéarité lorsque des modèles sont ajustés au moyen de données d'enquêtes.

Le facteur d'inflation de la variance (VIF, de l'anglais *variance inflation factors*) décrit à la section 2, qui représente l'une des techniques classiques de diagnostic de la colinéarité les plus répandues, s'applique principalement à des régressions par les moindres carrés ordinaires ou pondérées. Le VIF mesure l'augmentation (« inflation ») de la variance de l'estimation d'une pente causée par la non-orthogonalité des variables explicatives en sus de ce que la variance serait dans des conditions d'orthogonalité. À la section 3, nous examinons le cas d'un analyste qui estime les paramètres d'un modèle en appliquant la méthode des moindres carrés pondérés par les poids de sondage (MCPPS) et calculons les VIF appropriés pour cette méthode. Les composantes du VIF peuvent être estimées en utilisant les éléments d'un estimateur de variance d'usage fréquent dans les projections pour l'analyse des données d'enquêtes. Dans le cas de la régression linéaire, un estimateur de variance de type sandwich permettra d'estimer à la fois la variance sous le modèle et la variance sous le plan de sondage de l'estimateur MCPPS de la pente. Comme nous le montrerons à la section 3, la variance sous le modèle ou sous le plan de sondage de $\hat{\beta}_k$, un estimateur de la pente associée à la variable explicative \mathbf{x}_k , présente une certaine inflation quand diverses variables explicatives sont corréliées les unes aux autres, comparativement à ce qu'elle serait si \mathbf{x}_k était orthogonale aux autres variables explicatives. La mesure de l'inflation, le VIF, est composée de termes qui doivent être estimés d'après l'échantillon. Notre approche consiste à substituer des estimateurs qui ont une interprétation à la fois sous le modèle et sous le plan, comme il est décrit à la section 3.5.

À la quatrième section, nous présentons une étude empirique portant sur des données de la National Health and Nutrition Examination Survey réalisée aux États-Unis. Nous montrons l'application de notre nouvelle approche et comparons les valeurs nouvellement dérivées du VIF pour l'estimateur MCPPS à celles obtenues pour les estimateurs standard. Les comparaisons montrent que les valeurs du VIF diffèrent selon la méthode de régression et qu'un VIF spécial pour échantillon complexe devrait être utilisé pour évaluer l'effet nuisible de la colinéarité dans l'analyse des données d'enquêtes.

2. Diagnostics de colinéarité dans l'estimation par les moindres carrés ordinaires

Supposons que l'échantillon s contient n unités sur chacune desquelles sont observées p variables explicatives

Facteurs d'inflation de la variance dans l'analyse des données d'enquêtes complexes

Dan Liao et Richard Valliant

Résumé

Les données d'enquêtes servent souvent à ajuster des modèles de régression linéaire. Les valeurs des covariables utilisées dans ces modèles sont parfois mesurées à l'aveugle, c'est-à-dire sans connaissance de la réponse. Les valeurs des covariables utilisées dans les enquêtes sont souvent mesurées à l'aveugle, c'est-à-dire sans connaissance de la réponse. Les valeurs des covariables utilisées dans les enquêtes sont souvent mesurées à l'aveugle, c'est-à-dire sans connaissance de la réponse.

Mots clés : Echantillon en grappes ; diagnostics de colinéarité ; estimateur de la variance par linéarisation ; moindres carrés pondérés par les poids de sondage ; échantillon stratifié.

1. Introduction

Dans une régression linéaire, la colinéarité des variables

expliquent, en outre, que les variables sont affectées les unes aux autres. Les termes multicolinéarité et multicollinéarité sont également utilisés pour désigner ce genre de situation. La colinéarité est préoccu-

partie pour des raisons tant numériques que statistiques. Les estimations des coefficients de pente peuvent être numériquement instables dans certains jeux de données ce en sens que de petites variations dans les X ou les Y peuvent

produire de grandes variations dans les valeurs de ces estimations. Statistiquement, la corrélation entre les variables explicatives peut mener à des estimations de pente dont la variance est grande. En outre, quand les X sont

fortement corrélés, le R^2 d'une régression peut être grand alors que les estimations de pente individuelles ne sont pas statistiquement significatives. Même si elles le sont, elles peuvent être de signe opposé à celui attendu (Netter, Kutner,

Wasserman et Nachshheim (1996). La coliméarité influe parfois aussi sur les prévisions (Smith 1974 ; Belsley 1984). Dans les plans expérimentaux, il peut être possible de créer des situations où les variables explicatives sont ortho-

gonales les unes par rapport aux autres. Par contre, dans de nombreuses enquêtes, des données sur des variables fortement corrélées sont recueillies pour l'analyse. Ainsi, le revenu total et ses composantes (par exemple, traitements et

salaires, gains en capital, intérêts et dividendes) sont recueillis dans le cadre de la Panel Survey of Income Dynamics (<http://psidonline.isr.umich.edu/>) afin de suivre le

Alors que la littérature sur les diagnostics de régression (1986, section 10.3),

est abondante pour les données ne provenant pas d'enquêtes, elle l'est nettement moins pour les données d'enquêtes. Au cours de la dernière décennie, quelques articles ont présenté des techniques d'évaluation de la qualité de la

régression appliquée à des données d'enquêtes complexes, principalement pour repérer les points et les groupes influents présentant des valeurs de données ou de poids de sondage anormales. Ainsi, Elliott (2007) a élaboré des mé-

thodes bayésiennes de tronçonnage des séquences d'ADN. Les résultats obtenus sont comparés à ceux obtenus avec les méthodes classiques de tronçonnage. Les résultats obtenus sont comparés à ceux obtenus avec les méthodes classiques de tronçonnage.

Valliant (2009, 2011) ont adapté et étendu une série de techniques diagnostiques classiques à la régression sur des données d'enquêtes complexes, principalement pour la détection des observations influentes et des groupes d'obser-

ventions influents. Les travaux de recherche de Li portent sur les résidus et les leviers, DEBETA, DEFBETA, DFFIT, DFFITS, la distance de Cook et l'approche *forward search* (recherche avant). Alors que de nombreuses publications de

Qualité, L., et Tillé, Y. (2008). Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'Enquête suisse sur la valeur ajoutée. *Techniques d'enquête*, 34, 193-201.

Sändal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Smith, P., Pont, M. et Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994-2000. *The Statistician*, 52, 257-295.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, 288-289.

Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, 24, 53-78.

Lowerre, J.M. (1979). Sampling for change. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 343-347.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

Knothnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. New York : Springer-Verlag.

Knothnerus, P., et Van Deiden, A. (2006). Estimation of changes in repeated surveys and their significance, <http://www.user.essex.ac.uk/uisc/mojs2006/programme/data/paper/knothnerus.doc>.

Konschmink, C.A., Monsour, N.J. et Ditlefsen, R.E. (1985). Constructing and maintaining frames and samples for business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 113-122.

Lantel, N. (1987). Variances for a rotating sample from a changing population. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 496-500.

Lowerre, J.M. (1979). Sampling for change. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 343-347.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

$$\sigma_{yx} = \frac{E(\Delta y | \Delta x_0)}{E(\Delta x_0)} \sigma_x^2 \quad (18)$$

Donc, pour estimer, par exemple, $\text{cov}(n_{t-12}^{h,H+1}, n_{th}^{kh})$ dans (17) sous l'hypothèse de normalité, il suffit d'évaluer l'effet prévu sur $y = n_{th}^{kh}$ causé par une variation des futures unités

disparues $x = n_{th}^{h,H+1}$ dans s_{t-12}^h .

Soit $\Delta n_{t-12}^{h,H+1}$ une variation supplémentaire (positive) de ces unités disparues dans s_{t-12}^h . Définissons $p_{\text{janv},t}^{h,H+1}$ par

$p_{\text{janv},t}^{h,H+1} / N_{t-12}^{h,H+1} = N_{\text{janv},t}^{h,H+1} / N_{t-12}^{h,H+1}$ où $N_{\text{janv},t}^{h,H+1}$ est le nombre d'unités

disparues dans la strate h entre janvier et le mois t . En outre, $p_{t-12}^{hg} = N_{t-12}^{h,H+1} / N_{t-12}^h$ ($g = 1, \dots, H+1$). En utilisant l'hypothèse (v), le nombre prévu d'unités disparues supplémentaires dans l'échantillon de janvier avant le renouvel-

lement peut être estimé par $p_{\text{janv},t}^{h,H+1} \Delta n_{t-12}^{h,H+1}$. D'où, le nombre prévu d'unités disparues supplémentaires dans

l'échantillon après le renouvellement peut être estimé par

$$N_{\text{janv},t}^{\text{red}} p_{\text{janv},t}^{h,H+1} \Delta n_{t-12}^{h,H+1};$$

$$N_{\text{janv},t}^{\text{red}} = (0,9 - f_h) / (1 - f_h), \quad (19)$$

où $\gamma_{\text{janv},t}^{\text{red}}$ est le facteur de réduction dû au renouvellement de l'échantillon en janvier. Pour l'établissement de l'expression

(19), voir la fin de cette annexe. Les mises à jour mensuelles correspondantes entre janvier et le mois t dues à ces unités

disparues supplémentaires dans l'échantillon provenant de la strate h mènent à l'estimation suivante de l'augmentation

prévue du nombre d'unités *entrantes* n_{th}^{kh} provenant de la strate k ($k \neq h$) dans l'échantillon du mois t

$$E(\Delta n_{t-12}^{kh} | \Delta n_{\text{janv},t}^{h,H+1} \Delta n_{t-12}^{h,H+1} p_{\text{janv},t}^{h,H+1} p_{th}^{kh}), \quad (20)$$

où $p_{th}^{kh} = N_{t-12}^{kh} / (N_{t-12}^{h,H+1} - N_{t-12}^{hh})$. Rappelons, d'après la sous-section 2.1, qu'une mise à jour au mois s a lieu uniquement

quand $u_{s-1}^h \neq f^h D_{s-1}^h$, où $D_s^h(d_s^h)$ représente le nombre d'unités disparues dans $U_s^h(s_s^h)$, et que $n_t^{kh} = f^h N_{t-12,t}^{kh}$ est

fixé quand $N_{\text{janv},t+1}^{h,H+1} = 0$ ($k \neq h$) est remplacé. De surcroît, notons que les nouvelles unités sont exclues de la définition

de p_{th}^{kh} dans (20) à cause de l'hypothèse (iv).

Ensuite, définissons pour $m = 0, 1, 2$

$$\bar{O}_{t-m}^h = \frac{1}{N_{t-m}^h} \sum_{m=1}^{t-1} O_{t-m}^h;$$

$$(S_{t-m}^{h-m})^2 = \frac{1}{N_{t-m}^h} \sum_{m=1}^{t-1} (O_{t-m}^h)^2 - \bar{O}_{t-m}^h{}^2;$$

$$p_{t-12}^{h,H+1} = 1 - p_{t-12}^{h,h};$$

$$d_{t-12}^{h,h} = 1 - d_{t-12}^{h,H+1};$$

$$\bar{O}_{t-12}^{h,h} = \sum_{h \leq H} \bar{O}_{t-12}^{hg} \quad \bar{O}_{t-12}^{hg} = \sum_{h \leq H} \bar{d}_{t-12}^{hg} \frac{p_{t-12}^{kh}}{p_{t-12}^{hh}};$$

$$\bar{O}_{t-12}^{h,h} = \sum_{k \neq h} \bar{d}_{t-12}^{kh} \frac{p_{t-12}^{kh}}{p_{t-12}^{hh}}.$$

que

$$\left| C_{h,h}^{\text{sec}} \right| \leq \frac{N_{\text{janv},t}^{\text{red}} p_{\text{janv},t}^{h,H+1} p_{t-12}^{h,h} (1 - p_{t-12}^{h,H+1}) (1 - f_h) (S_t^h)^2}{N_{t-12}^{hh}}.$$

En supposant que les deux termes entre parenthèses dans (24) sont absolument plus petits que S_t^h , il découle de (24)

$$C_{h,h}^{\text{sec}} = A_h^h (\bar{O}_{t-12}^{h,H+1} - \bar{O}_{t-12}^{hh}) / n_{th}^{hh}. \quad (24)$$

respectivement ($1 \leq g \leq H$). Maintenant, en substituant

$$\text{acov}(n_{t-12}^{hg}, n_{th}^{hh}) = n_{t-12}^{hg} p_{t-12}^{hg} A_h^h / p_{th}^{hh}, \quad \text{acov}(n_{t-12}^{hh}, n_{th}^{hh}) = -n_{t-12}^{hh} A_h^h;$$

même, pour $k = h$, nous obtenons d'après (21) et (22b)

$$\left\{ \begin{array}{l} \delta_{t-12}^{hg} = \text{cov}(n_{t-12}^{hg}, n_{th}^{hh}) - \text{cov}(n_{t-12}^{h,H+1}, n_{th}^{hh}) / N_{t-12,t}^{h,H+1} \text{ d'où découle l'expression (22a)} \\ \delta_{t-12}^{hg} = 0 \text{ autrement} \end{array} \right.$$

où

$$\text{cov}(n_{t-12}^{h,H+1}, n_{th}^{hh}) = -\text{cov}(n_{t-12}^{h,h}, n_{th}^{hh}) = -\sum_{h \leq H} \sum_{t \in U_{t-12}^{hg}} \text{cov}(\delta_{t-12}^{hg}, n_{th}^{hh}),$$

où (21) est utilisée également. Ou bien, notons que

$$= -n_{t-12}^{h,h} \frac{p_{t-12}^{hg}}{p_{t-12}^{hh}} A_h^h \quad (22b)$$

$$\text{acov}(n_{t-12}^{hg}, n_{th}^{hh}) = -\frac{p_{t-12}^{hg}}{p_{t-12}^{hh}} \text{acov}(n_{t-12}^{h,H+1}, n_{th}^{hh}) \quad (22a)$$

relation suivante comme approximation de $\text{cov}(n_{t-12}^{hg}, n_{th}^{hh})$

où, pour simplifier, nous avons omis le terme $N_{t-12}^{h,H+1} / (N_{t-12}^h - 1)$ à la deuxième ligne. Comme $n_{t-12}^{h,H+1}$ est fixé, l'expression $\text{cov}(n_{t-12}^{h,H+1}, n_{th}^{hh}) = -\text{cov}(n_{t-12}^{h,h}, n_{th}^{hh})$ est vérifiée. D'où, par analogie avec la loi multihypergéométrique, nous pouvons utiliser pour $1 \leq g \leq H$ et $k \neq h$ la

$$\begin{aligned} A_h^h &= \gamma_{\text{janv},t}^{\text{red}} p_{\text{janv},t}^{h,H+1} p_{t-12}^{h,H+1} (1 - p_{t-12}^{h,H+1}) (1 - f_h), \\ &\approx n_{th}^{hh} A_h^h / p_{th}^{hh}; \\ &\approx \gamma_{\text{janv},t}^{\text{red}} p_{\text{janv},t}^{h,H+1} p_{t-12}^{h,H+1} p_{t-12}^{h,h} n_{th}^{hh} / (1 - p_{t-12}^{h,H+1}) (1 - f_h), \\ &= \frac{E(\Delta n_{t-12}^{kh} | \Delta n_{t-12}^{h,H+1})}{\text{var}(n_{t-12}^{h,H+1})} \text{var}(n_{t-12}^{kh}) \end{aligned} \quad (21)$$

Maintenant, en nous servant de (18) et (20), nous obtenons pour $k \neq h$ l'approximation de la covariance suivante

Donc, quand $p_{t-12}^{h,h}, p_{t-12}^{h,H+1} \leq 0,1$, nous pouvons conclure que, sous les hypothèses susmentionnées, la contribution de la seconde composante de la covariance est inférieure à 1 %

Premièrement, considérons le cas des strates sans nouvelles unités et sans unités disparues. À part la mise à jour annuelle de janvier, il n'y a dans ces conditions aucune mise à jour mensuelle. D'où, $n_{h,t}^{\text{dec,janv}} = n_{h,t}^{\text{dec,REQ}}$ est fixe et il en découle (5). Ce cas s'applique aux supermarchés assez stable au fil des ans. Deuxièmement, dans le cas des nouvelles unités et des unités disparues dans les strates, nous pouvons écrire $n_{h,t}^{\text{dec}}$ sous la forme

Justification de (5)

Annexe A

Les opinions exprimées dans le présent article sont celles de Statistics Netherlands. Les auteurs remercient le rédacteur associé et deux examinateurs anonymes de leurs commentaires et suggestions constructifs qui leur ont permis d'améliorer considérablement l'article.

Remerciements

L'exemple des supermarchés des Pays-Bas illustre l'une des applications pratiques des formules de variance, à savoir déterminer quel estimateur possède la variance la plus faible. Les résultats confirment que la variance de l'estimateur simple $\hat{g}_{h,t-13}$ est proche de celle de l'estimateur $\hat{g}_{h,t-13}^{\text{réel}}$ de la section 4 qui comprend une correction pour tenir compte du renouvellement de l'échantillon en janvier. Donc, pour les supermarchés des Pays-Bas, $\text{var}(\hat{g}_{h,t-13}^{\text{réel}})$ pourrait être utilisé pour estimer $\text{var}(\hat{g}_{h,t-13}^{\text{réel}})$. Pour des branches d'activité ayant un autre code de la CTI, il convient de vérifier si $\text{var}(\hat{g}_{h,t-13}^{\text{réel}}) \approx \text{var}(\hat{g}_{h,t-13})$, puisque l'effet du renouvellement de l'échantillon en janvier ne doit pas nécessairement être négligeable.

Pour le plan d'échantillonnage des supermarchés des Pays-Bas, le deuxième terme de covariance dans (6) est négligeable parce que $n_{h,t}^{\text{dec,janv}}$ est fixé. Par contre, pour le plan d'échantillonnage SAMU utilisé en Suède, ce terme est non négligeable et son estimation prend beaucoup de temps; le terme SAMU (SAMordnade Urväl) est un acronyme suédois signifiant échantillons coordonnés. À l'annexe A, nous proposons une méthode de rechange pour estimer cette covariance. Cependant, sous la condition que $\hat{g}_{h,t-12}^{\text{ps, sous}} \approx \hat{g}_{h,t-12}^{\text{ps, sous}}$ il suffit à notre avis d'utiliser uniquement la première covariance. Cela simplifie considérablement la procédure d'estimation. En outre, sous l'hypothèse de normalité, l'intervalle de confiance conditionnel possède de meilleures propriétés de couverture que l'intervalle conditionnel.

De plus, pour une variation donnée Δx_0 de x_t , l'espérance conditionnelle de la variation de y est égale à $E(\Delta y | \Delta x_0) = \sigma_{yx}^2 / \sigma_x^2$ ou, de manière équivalente, à

$$E(y | x_0) = \mu_y + \frac{\sigma_{yx}^2}{\sigma_x^2} (x_0 - \mu_x).$$

Afin d'estimer les covariances dans (17), considérons la hypothèses susmentionnées, $C_{h,t,\text{sec}} = 0$ pour $\ell \neq h$. Notons que, sous les

$$C_{h,t,\text{sec}} \equiv \frac{1}{n_{h,t-12}^h} \left\{ E \left(\sum_{k=1}^{H+1} m_{h,t-12}^{hg} v_h \right) \text{cov} \left(\sum_{k=0}^H m_{h,t-12}^{kh} v_h \right) \right\} = \frac{1}{n_{h,t-12}^h} \sum_{k=1}^{H+1} \sum_{\ell=1}^{H+1} \bar{O}_{h,t-12}^{\ell k} \text{cov}(n_{h,t-12}^{hg}, n_{h,t-12}^{k\ell}). \quad (17)$$

En fait, nous avons supposé jusqu'à présent que la distribution de $n_{h,t}^{k\ell}$ ($k \neq h$) peut être décrite par une loi hypergéométrique ayant les paramètres $(N_t^{k\ell}, N_{t-12}^{k\ell}, n_t^{k\ell})$. Une remarque similaire s'applique à $n_{h,t-12}^{0h}$. Cependant, nous pouvons soutenir qu'en pratique, ces hypothèses donnent lieu à une erreur mineure, de deuxième ordre, dans les formules de variance. Afin de dépister cette erreur, nous supposons pour simplifier et sans perte de généralité que i) les nouvelles unités et les unités disparues ne migrent pas entre les strates, ii) il n'y a pas de disparitions parmi les nouvelles unités, iii) $n_t^{0h} = f_h N_{0h}^t$ est fixe, iv) après leur premier mois dans la population, les nouvelles unités sont sans pertinence pour les mises à jour mensuelles durant le reste de la période étudiée et v) les unités disparues ne sont pas sélectionnées dans l'échantillon ni supplantées de celui-ci lors des mises à jour mensuelles; donc, une erreur de troisième ordre demeure ignorée. Sous ces hypothèses, nous examinons maintenant de plus près la seconde composante de la covariance pour $\ell = h$, disons $C_{h,t,\text{sec}}^h$ provenant de (4a). Par analogie avec (6), $C_{h,t,\text{sec}}^h$ peut s'écrire

$$n_{h,t}^{\ell\ell} = n_{h,t-12}^{\ell\ell} - \sum_{k \neq h}^H n_{h,t-12}^{k\ell} \quad (16)$$

Knothnerus et van Delden : À propos de la variance des variations estimées

de Nordberg (3.9) pour la covariance entre O_{t-13}^i et O_t^i est égal à 111,1 millions, tandis que l'estimateur de covariance (13) proposé donne 67,8 millions.

Tableau 2
Données de panel¹ provenant d'une population avec $N = 50$ et $H = 1$

Chiffre d'affaires par unité (en milliers d'euros)	1	2	3	4	5
$t = 3$	493,9	264,3	1 179,1	380,0	
$t = 16$	475,3	472,0	267,0	1 169,0	

¹ En réalité, les données de panel appartiennent à la sous-strate $h_t^i = 65$.

5. Conclusion

Les formules de variance obtenues dans le présent article sont utiles pour calculer la variance du taux de croissance annuel estimé du chiffre d'affaires mensuel. L'utilisation de (13) comme estimateur de $\text{cov}(O_{t-12}^i, O_t^i)$ produit, en particulier, des estimations raisonnables de la covariance des variations. La procédure d'estimation de la variance tient compte de l'utilisation de panels rotatifs, des nouvelles unités, des unités disparues et des unités qui migrent entre les strates.

En outre, nous recommandons d'estimer une covariance de population au moyen de l'estimateur (12) fondé sur la corrélation correspondante estimée d'après l'échantillon chevauchant, ainsi que sur les variances correspondantes estimées d'après de grands échantillons distincts. Cela peut aider à éviter une sous-estimation importante ou la production d'un résultat négatif par l'estimateur de la variance du taux de croissance annuel. Les covariances estimées résultantes ne présentent qu'un léger biais.

Comme il est décrit à la section 3, nous avons utilisé la corrélation estimée $\hat{\rho}_{h_t^i, \text{CHEV}}^{t-13, t}$ provenant de la partie chevauchante de l'échantillon $s_{h_t^i, \text{CHEV}}^{t-13, t}$ pour estimer la covariance $S_{h_t^i, \text{CHEV}}^{t-13, t}$ afin d'éviter que (11) donne des résultats négatifs. Knothnerus et Van Delden (2006) ont évalué le biais de $\hat{\rho}_{h_t^i, \text{CHEV}}^{t-13, t}$ pour les données sur les supermarchés des Pays-Bas et découvert une faible sous-estimation de $\hat{\rho}_{h_t^i, \text{CHEV}}^{t-13, t}$ résultant d'une légère surestimation, inférieure à 5 %, de $\text{var}(\hat{\rho}_{h_t^i, \text{CHEV}}^{t-13, t})$.

L'utilisation de l'estimateur $S_{h_t^i, \text{CHEV}}^{t-13, t}$ dans (10) peut donner un résultat négatif pour (11) et une corrélation estimée $\hat{\rho}_{h_t^i, \text{CHEV}}^{t-13, t}$ plus grande que 1. Par exemple, considérons la population spécifique pour laquelle $N = 50$ et $H = 1$

composée des unités de la sous-strate $h_t^i = 65$. Au moyen des données de panel pour cette population, présentées au tableau 2 pour $t = 3$ et $t = 16$, nous obtenons après certains calculs $S_{h_t^i, \text{NBG}}^{t-13, t} = 410,7$, $S_t^i = 394,3$ et $G_{t-13}^i = 1,028$. Notons, dans la suite de la présente section, l'indice $h_t^i = 11$ est omis dans les symboles parce qu'il n'existe qu'une seule strate. Le tableau 3 donne, pour trois approches différentes, certaines estimations supplémentaires pour les données de panel du tableau 2. Par exemple, l'utilisation de $S_{h_t^i, \text{CHEV}}^{t-13, t}$ dans (10) produit une corrélation estimée $\hat{\rho}_{h_t^i, \text{CHEV}}^{t-13, t} = 1,39$. Celle-ci donne alors, en utilisant (11), une estimation négative de la variance égale à moins 2,2 millions. Similairement, pour les mêmes données, l'estimateur de $S_{h_t^i, \text{NBG}}^{t-13, t}$ de $S_{h_t^i, \text{CHEV}}^{t-13, t}$ fondé sur Nordberg (2000) produit pour (11) le résultat moins 36,9 millions, parce que l'estimation correspondante $\hat{\rho}_{h_t^i, \text{NBG}}^{t-13, t}$ devient 1,64. En revanche, l'utilisation de la corrélation estimée d'après l'échantillon chevauchant $s_{h_t^i, \text{CHEV}}^{t-13, t}$ conformément à (12) donne $\hat{\rho}_{h_t^i, \text{CHEV}}^{t-13, t} = 0,9997$ et l'estimation de la variance au moyen de (11) devient la valeur positive de 52,1 millions. En outre, pour les données de panel du tableau 2, le résultat de l'estimateur

Tableau 3
Estimations selon trois approches différentes

Approche					Paramètres à estimer	
Nordberg(2000)	estimateur	$S_{t-13,t}^{NBG}$	$S_{t-13,t}^{NBG}$	$S_{t-13,t}^{NBG}$	$\hat{S}_{t-13,t}^{NBG}$	
	résultat	$265,2 \times 10^3$	$\frac{\hat{S}_{t-13,t}^{NBG}}{\hat{S}_{t-13,t}^{CHEV}}$	$1,64$	$\hat{S}_{t-13,t}^{NBG}$	
Eq. (10)	estimateur	$S_{t-1,t}^{CHEV}$	$\frac{\hat{S}_{t-13,t}^{CHEV}}{\hat{S}_{t-13,t}^{CHEV}}$	$1,39$	$\hat{S}_{t-13,t}^{CHEV}$	
	résultat	$225,0 \times 10^3$	$\hat{S}_{t-13,t}^{CHEV}$	$1,00$	$\hat{S}_{t-13,t}^{CHEV}$	
Eq. (12)	estimateur	$\hat{\rho}_{t-13,t}^{CHEV}$	$\frac{\hat{\rho}_{t-13,t}^{CHEV}}{\hat{\rho}_{t-13,t}^{CHEV}}$	$1,00$	$\hat{\rho}_{t-13,t}^{CHEV}$	
	résultat	$161,9 \times 10^3$	$\hat{\rho}_{t-13,t}^{CHEV}$	$1,00$	$\hat{\rho}_{t-13,t}^{CHEV}$	
En fait, 0,9997.						

Taux de croissance estimés avec les marges d'erreur à 95 %		
t	$\hat{g}_{t,t-13}^{réel} \times 100\%$	$\hat{g}_{t,t-13} \times 100\%$
16	-0,3 (± 1,0)	-0,4 (± 1,0)
17	-3,7 (± 1,0)	-3,8 (± 0,9)
18	1,6 (± 1,0)	1,5 (± 0,9)
19	-2,2 (± 0,9)	-2,3 (± 0,9)
20	0,5 (± 0,8)	0,4 (± 0,7)
21	-1,7 (± 0,8)	-1,8 (± 0,7)
22	-2,2 (± 0,8)	-2,3 (± 0,7)
23	0,0 (± 0,8)	-0,1 (± 0,7)
24	-2,5 (± 0,9)	-2,4 (± 0,9)

Les marges d'erreur à 95 % sont données entre parenthèses.

Le tableau 1 donne les taux de croissance annuels et leur marge d'erreur à 95 % pour $t = 16, \dots, 24$. Nous constatons que les marges d'erreur à 95 % des taux de croissance estimés $\hat{g}_{t,t-13}^{réel}$, utilisées à l'heure actuelle par Statistics Netherlands, varient entre 0,8 et 1,0 (point de pourcentage). Par exemple, à la première période ($t = 16$), l'intervalle de confiance à 95 % du taux de croissance annuel va de -1,3 à 0,7 %. Comme prévu, les marges d'erreur à 95 % pour l'estimateur plus compliqué $\hat{g}_{t,t-13}^{réel}$ sont proches de celles pour l'estimateur plus simple $\hat{g}_{t,t-13}$ provenant de (1). Les marges d'erreur à 95 % de $\hat{g}_{t,t-13}$ varient entre 0,7 et 1,0 (point de pourcentage). L'estimateur du taux de croissance (point de pourcentage). L'estimateur du taux de croissance qu'il convient de privilégier est $\hat{g}_{t,t-13}^{réel}$ car il comprend une correction pour la mise à jour annuelle de l'échantillon en janvier. L'estimation de sa variance peut toutefois être simplifiée en utilisant l'estimateur de la variance décrit à la section 3, plutôt que l'expression plus laborieuse pour $\text{Var}(\hat{g}_{t,t-13}^{réel})$.

4.3 Résultats

Le tableau 1 donne les taux de croissance annuels et leur marge d'erreur à 95 % pour $t = 16, \dots, 24$. Nous constatons que les marges d'erreur à 95 % des taux de croissance estimés $\hat{g}_{t,t-13}^{réel}$, utilisées à l'heure actuelle par Statistics Netherlands, varient entre 0,8 et 1,0 (point de pourcentage). Par exemple, à la première période ($t = 16$), l'intervalle de confiance à 95 % du taux de croissance annuel va de -1,3 à 0,7 %. Comme prévu, les marges d'erreur à 95 % pour l'estimateur plus compliqué $\hat{g}_{t,t-13}^{réel}$ sont proches de celles pour l'estimateur plus simple $\hat{g}_{t,t-13}$ provenant de (1). Les marges d'erreur à 95 % de $\hat{g}_{t,t-13}$ varient entre 0,7 et 1,0 (point de pourcentage). L'estimateur du taux de croissance (point de pourcentage). L'estimateur du taux de croissance qu'il convient de privilégier est $\hat{g}_{t,t-13}^{réel}$ car il comprend une correction pour la mise à jour annuelle de l'échantillon en janvier. L'estimation de sa variance peut toutefois être simplifiée en utilisant l'estimateur de la variance décrit à la section 3, plutôt que l'expression plus laborieuse pour $\text{Var}(\hat{g}_{t,t-13}^{réel})$.

Dans le présent exemple, les calculs des variances et des intervalles de confiance sont fondés sur les données sur le chiffre d'affaires des supermarchés des Pays-Bas pour des périodes de quatre semaines en 2003 et en 2004 (c'est-à-dire $t = 1, \dots, 26$). Donc, il existe 13 observations pour une année et, conséquemment, nous utilisons des symboles légèrement adaptés, tels que $\hat{g}_{t,t-13}$ dans la suite de la présente section.

La population est constituée d'environ 3 500 établissements. Les données sur le chiffre d'affaires proviennent d'un échantillon stratifié et de fichiers administratifs. Un échantillon EASS brut d'environ 900 unités stratifiées selon la taille est tiré de la liste complète des unités de population

4.2 Description des données

un développement en série de Taylor de premier ordre.

À la présente section, nous allons comparer les variances des estimateurs (1) et (15). De la même façon que (2), les formules de variance pour $\hat{g}_{t,t-12}^{réel}$ peuvent être établies par

$$\hat{O}_t^{réel} = 1 + \hat{g}_{t,t-12}^{réel} = \prod_{j=0}^{t-1} (1 + \hat{g}_{t-j,t-j-1})$$
$$\hat{O}_t = \hat{O}_{t-1} \times \frac{\hat{O}_{t-1}}{\hat{O}_{t-2}} \times \dots \times \frac{\hat{O}_{t-1}}{\hat{O}_{t-12}} \times \frac{\hat{O}_{t-1}}{\hat{O}_{janvA}} \times \frac{\hat{O}_{t-1}}{\hat{O}_{janvN}} \quad (15)$$

Contrairement à l'estimateur (1), l'estimateur réel utilise le chiffre d'affaires mensuel est fondé sur une chaîne de 12 variations mensuelles du chiffre d'affaires.

En janvier, le chiffre d'affaires mensuel est fondé sur une chaîne de 12 variations mensuelles du chiffre d'affaires. Contrairement à l'estimateur (1), l'estimateur réel utilise le chiffre d'affaires mensuel est fondé sur une chaîne de 12 variations mensuelles du chiffre d'affaires. Contrairement à l'estimateur (1), l'estimateur réel utilise le chiffre d'affaires mensuel est fondé sur une chaîne de 12 variations mensuelles du chiffre d'affaires.

4.1 Deux estimateurs de la variation annuelle du chiffre d'affaires

4. Une application à la variation du chiffre d'affaires des supermarchés des Pays-Bas

indispensable, y compris l'estimation de la deuxième composante selon Nordberg. Pour une approche différente du problème d'estimation de la deuxième composante, voir l'annexe A.

Pour une justification de l'utilisation d'une (co)variance conditionnelle, voir Holt et Smith (1979). Un avantage important de la (co)variance conditionnelle est que l'intervalle de confiance correspondant possède de meilleures propriétés de couverture que celui fondé sur la variance inconditionnelle. Désignons l'intervalle de confiance à 95 % conditionnel classique pour un paramètre arbitraire θ par $(\hat{\theta}, \hat{\theta}^u | v)$, où v désigne le vecteur constitué de toutes les statistiques (auxiliaires) qui interviennent dans les (co)variances conditionnelles. Alors, sous l'hypothèse de normalité et certaines conditions faibles, il est vérifié que le niveau de confiance (NC) à 95 % réel est égal au niveau de confiance nominal

parce que

$$NC = \sum_{v \in \mathcal{V}} P(v) P(\hat{\theta} < \theta < \hat{\theta}^u | v) = \sum_{v \in \mathcal{V}} P(v) = 0,95,$$

où \mathcal{V} représente l'ensemble de tous les résultats possibles du vecteur aléatoire v . Lorsqu'on utilise des (co)variances inconditionnelles, les intervalles de confiance obtenus peuvent être assez inexacts pour une répartition donnée de l'échantillon. En outre, quand la moyenne est calculée sur l'ensemble des répartitions, NC peut différer de 0,95 ; par exemple, voir Knothnerus (2003, pages 133-135). Notons qu'à l'étape de la planification avant le tirage de l'échantillon, les variances inconditionnelles sont toujours utiles pour examiner l'effet de plan de Kish en vue de comparer divers plans d'échantillonnage. De surcroît, notons que, pour évaluer un intervalle de confiance conditionnel pour $\sigma_{i-12,i}^2$, les variances sous-jacentes de $O_{i-12,i}^{ps, sous}$ devraient également être prises conditionnellement à v^{hi} ($m = 0, 12$).

Enfin, l'estimateur sans biais proposé par Nordberg (2000, équation (3.9)) pour la première composante de (6) est assez différent de ceux décrits à la sous-section précédente. En fait, son estimateur est fondé sur la procédure d'estimation du terme de covariance $S_{i-12,i}^{ht}$ qui suit. Premièrement, estimer la quantité sous-jacente $\sum_{i=1}^{N_{i-12,i}} O_{i-12,i}^{ht}$ à partir de l'échantillon chevauchant $s_{i-12,i}^{ht}$. Deuxièmement, estimer les moyennes du chiffre d'affaires correspondantes d'après $s_{i-12,i}^{ht}$ et s_i^{ht} respectivement. Puisque les proportions estimées de cette façon proviennent d'échantillons différents, il n'est pas toujours possible d'éviter que (11) donne un résultat négatif. Pour un petit exemple portant sur des données réelles, voir la section qui suit. Dans la suite de l'exposé, l'estimateur sous-jacent de Nordberg pour $S_{i-12,i}^{ht}$ est désigné par $\hat{S}_{i-12,i}^{htNBG}$. La détermination de l'expression explicite pour $\hat{S}_{i-12,i}^{htNBG}$ peut être obtenue sur demande auprès des auteurs.

3.4 Une comparaison avec les résultats de Nordberg

En utilisant la formulation classique des indicateurs d'inclusion δ_i^{ht} pour chaque strate, Nordberg (2000) dérive une expression différente pour la première composante de l'expression (3.4) ; une preuve peut être obtenue sur demande auprès des auteurs. En outre, Nordberg établit une expression non nulle pour la deuxième composante de (6), c'est-à-dire la covariance entre les deux espérances conditionnelles de la Suède diffère légèrement du nôtre.

Selon Nordberg (2000, page 370), l'estimation de la deuxième composante pour le plan d'échantillonnage de la Suède nécessite une procédure gourmande en ressources informatiques qui comprend la simulation du procédé d'échantillonnage. Cependant, puisque les $n_{i-12,i}^{ht}$ et $n_{i-12,i}^{ht}$ sont tous des statistiques auxiliaires, une autre option pourrait consister à conditionner sur ces statistiques afin de pouvoir ignorer la deuxième composante. Rappelons qu'une statistique est dite auxiliaire quand sa distribution marginale ne dépend pas des paramètres cibles à estimer ; voir Cox et Hinkley (1974, pages 31-35). Une approche de rechange de ce genre sans la deuxième composante est recommandée en particulier quand $\hat{\sigma}_{i-12,i}^{E, sous} \approx \hat{\sigma}_{i-12,i}^{ps, sous}$ ou $\hat{\sigma}_{i-12,i}^{ps, sous}$ est l'estimateur poststratifié basé sur les sous-strates h_i . Cependant, quand la différence entre $\hat{\sigma}_{i-12,i}^{E, sous}$ et $\hat{\sigma}_{i-12,i}^{ps, sous}$ est non négligeable, le calcul de la variance inconditionnelle paraît

$$\begin{aligned} & \text{provenant de (13) est égale à} \\ (14) \quad & \left[E \left\{ n_{i-12,i}^{ht} \left(1 - \frac{n_{i-12,i}^{ht} N_{i-12,i}^{ht}}{n_{i-12,i}^{ht} N_{i-12,i}^{ht}} \right) \right\} - \frac{n_{i-12,i}^{ht} N_{i-12,i}^{ht}}{n_{i-12,i}^{ht} N_{i-12,i}^{ht}} \right] S_{i-12,i}^{ht} \\ & = \left[E \left\{ n_{i-12,i}^{ht} \left(1 - \frac{n_{i-12,i}^{ht} N_{i-12,i}^{ht}}{n_{i-12,i}^{ht} N_{i-12,i}^{ht}} \right) \right\} - \frac{n_{i-12,i}^{ht} N_{i-12,i}^{ht}}{n_{i-12,i}^{ht} N_{i-12,i}^{ht}} \right] S_{i-12,i}^{ht} \end{aligned}$$

de cette hypothèse, un terme de ce genre avec $n_{i-12,i}^{ht} = 0$ ($m = 0$ ou $m = 12$) peut être négligé parce que l'espérance des $S_{i-12,i}^{ht}$ restants soient estimés de façon non biaisée. Sous négligé sans affecter son absence de biais, à condition que le terme de covariance correspondant dans (13) peut être En outre, notons que, si $n_{i-12,i}^{ht} = 0$ ($m = 0$ ou $m = 12$),

Nous proposons pour $S_{h_{t-12,t}}^{h_{t-12,t}}$ l'estimateur modifié suivant

$$(12) \quad \hat{S}_{h_{t-12,t}}^{h_{t-12,t}} = \hat{p}_{h_{t-12,t}}^{h_{t-12,t}} \hat{S}_{h_{t-12,t}}^{h_{t-12,t}} \hat{S}_{h_{t-12,t}}^{h_{t-12,t}}$$

où $\hat{p}_{h_{t-12,t}}^{h_{t-12,t}}$ est la corrélation entre les variables o^t et o^{t-12} dans $U_{h_{t-12,t}}^{h_{t-12,t}}$, et $\hat{p}_{h_{t-12,t}}^{h_{t-12,t}}$ est son estimation d'après $s_{h_{t-12,t}}^{h_{t-12,t}}$. Selon (10) et (12), la covariance (3) peut être estimée par

$$(13) \quad \text{cov}(\hat{O}^{t-12}, \hat{O}^t) = \sum_H \sum_{h_{t-12,t}} \frac{N_{h_{t-12,t}}^h N_{h_{t-12,t}}^t}{N_{h_{t-12,t}}^h N_{h_{t-12,t}}^t} \left(1 - \frac{N_{h_{t-12,t}}^h N_{h_{t-12,t}}^t}{N_{h_{t-12,t}}^h N_{h_{t-12,t}}^t} \right) \hat{S}_{h_{t-12,t}}^{h_{t-12,t}}.$$

Pour l'estimation $\hat{p}_{h_{t-12,t}}^{h_{t-12,t}}$, l'expression $\hat{p}_{h_{t-12,t}}^{h_{t-12,t}} \leq 1$ est toujours vérifiée, tandis que l'utilisation de (10) peut mener implicitement à une corrélation estimée plus grande que 1 et un résultat de (11) éventuellement négatif. Voir la section suivante pour un exemple. Dans toutes les applications rencontrées jusqu'à présent, les résultats négatifs de (11) pouvaient être expliqués par le fait que, contrairement à (12), l'utilisation de (10) mène implicitement à une corrélation estimée supérieure à 1. Cela concorde avec la constatation de Berger (2004, page 462) qu'une surestimation de la corrélation entre O^{t-12} et O^t peut donner lieu à une sous-estimation importante de la variance d'une variation. Néanmoins, dans certaines circonstances extraordinaires, l'utilisation de (12) pourrait produire un résultat négatif de (11) également. Des conditions suffisantes pour que l'utilisation de (12) produise un estimateur de variance non négatif avec une probabilité de 1 peuvent être obtenues sur demande auprès des auteurs. Pour une revue générale des méthodes d'estimation de la variance dans les enquêtes-entreprises, voir Brodie (2003).

L'application de (12) peut donner lieu à un problème particulier quand $n_{h_{t-12,t}}^{h_{t-12,t}} = 1$ ou $n_{h_{t-12,t}}^{h_{t-12,t}} = 1$. Pour évaluer les variances d'échantillon requises, on peut emprunter la variance d'échantillon durant un mois antérieur. On peut aussi imputer une variance quand il se dégage des données une relation de la forme $S_{h_{t-12,t}}^2 \approx c^2 \hat{O}_{h_{t-12,t}}^2$; voir Särnidal, Swensson et Wretman (1992, page 461). En outre, le terme de covariance correspondant pourrait être ignoré quand sa contribution (prévue) à la variance totale est faible. Cela est souvent le cas quand les fractions d'échantillonnage dans les strates h et ℓ sont faibles, c'est-à-dire dans les strates contenant des unités assez petites et, par conséquent, ayant de faibles variances comparativement aux strates contenant des unités plus grandes. Des remarques similaires s'appliquent aux $p_{h_{t-12,t}}^{h_{t-12,t}}$ imputées quand $n_{h_{t-12,t}}^{h_{t-12,t}} \leq 2$ et $n_{h_{t-12,t}}^{h_{t-12,t}} \geq 2$ ($m = 0, 1, 2$). Puisque les $p_{h_{t-12,t}}^{h_{t-12,t}}$ sont souvent assez élevées, cela semble être un moyen viable. Dans l'exemple donné à la section 4, les $p_{h_{t-12,t}}^{h_{t-12,t}}$ ont une moyenne globale de 0,90 et une variance de 0,0074, de sorte que l'effet des $p_{h_{t-12,t}}^{h_{t-12,t}}$ imputées sur les résultats finaux est vraisemblablement modéré.

À la dernière ligne, nous avons utilisé l'expression (26) de l'annexe B. En outre,

$$(7) \quad E\{\text{cov}(n_{h_{t-12,t}}^{h_{t-12,t}} \hat{O}_{h_{t-12,t}}^{h_{t-12,t}} | v_{h_{t-12,t}}^{h_{t-12,t}})\} = E\{n_{h_{t-12,t}}^{h_{t-12,t}} \hat{O}_{h_{t-12,t}}^{h_{t-12,t}} | v_{h_{t-12,t}}^{h_{t-12,t}}\} = E\left\{n_{h_{t-12,t}}^{h_{t-12,t}} \left(\frac{N_{h_{t-12,t}}^{h_{t-12,t}}}{N_{h_{t-12,t}}^{h_{t-12,t}}} - \frac{N_{h_{t-12,t}}^{h_{t-12,t}}}{N_{h_{t-12,t}}^{h_{t-12,t}}} \right) S_{h_{t-12,t}}^{h_{t-12,t}} \right\}.$$

La deuxième composante du deuxième membre de (6) est égale à $\hat{O}_{h_{t-12,t}}^{h_{t-12,t}} \text{cov}(n_{h_{t-12,t}}^{h_{t-12,t}}, n_{h_{t-12,t}}^{h_{t-12,t}}) = 0$ compte tenu de (5). Il découle donc de (4) et de (6) que

$$(8) \quad S_{h_{t-12,t}}^{h_{t-12,t}} = \frac{1}{N_{h_{t-12,t}}^{h_{t-12,t}}} \sum_{h_{t-12,t}} (O_{h_{t-12,t}}^{h_{t-12,t}} - \bar{O}_{h_{t-12,t}}^{h_{t-12,t}})(O_{h_{t-12,t}}^{h_{t-12,t}} - \bar{O}_{h_{t-12,t}}^{h_{t-12,t}}).$$

l'expression (9) peut être estimée d'après l'échantillon chevauchant $s_{h_{t-12,t}}^{h_{t-12,t}}$ par

$$(10) \quad \text{cov}(\hat{O}_{h_{t-12,t}}^{h_{t-12,t}}, \hat{O}_{h_{t-12,t}}^{h_{t-12,t}}) = \frac{n_{h_{t-12,t}}^{h_{t-12,t}} n_{h_{t-12,t}}^{h_{t-12,t}}}{N_{h_{t-12,t}}^{h_{t-12,t}}} \left(\frac{1}{N_{h_{t-12,t}}^{h_{t-12,t}}} - \frac{N_{h_{t-12,t}}^{h_{t-12,t}}}{N_{h_{t-12,t}}^{h_{t-12,t}}} \right) \hat{S}_{h_{t-12,t}}^{h_{t-12,t}},$$

Notons que (10) est sans biais pour l'estimation de (9) car

$$E(S_{h_{t-12,t}}^{h_{t-12,t}} | v_{h_{t-12,t}}^{h_{t-12,t}}) = S_{h_{t-12,t}}^{h_{t-12,t}}.$$

Bien que (10) produise des estimations raisonnables si $n_{h_{t-12,t}}^{h_{t-12,t}}$ est suffisamment grand, un inconvénient de l'estimateur de la covariance $\hat{S}_{h_{t-12,t}}^{h_{t-12,t}}$ dans (10) est que, si $n_{h_{t-12,t}}^{h_{t-12,t}}$ est petit, l'estimation de $\text{var}(\hat{O}_{h_{t-12,t}}^{h_{t-12,t}})$ dans le numérateur de (2) peut être négative. Rappelons que cette variance est estimée par

$$(11) \quad \text{var}(\hat{O}_{h_{t-12,t}}^{h_{t-12,t}}) = \text{var}(\hat{O}_{h_{t-12,t}}^{h_{t-12,t}}) + (\hat{O}_{h_{t-12,t}}^{h_{t-12,t}} - \bar{O}_{h_{t-12,t}}^{h_{t-12,t}})^2 \text{var}(\hat{O}_{h_{t-12,t}}^{h_{t-12,t}}) - 2\hat{O}_{h_{t-12,t}}^{h_{t-12,t}} \text{cov}(\hat{O}_{h_{t-12,t}}^{h_{t-12,t}}, \hat{O}_{h_{t-12,t}}^{h_{t-12,t}}).$$

Donc, nous proposons un estimateur pour remplacer $\hat{S}_{h_{t-12,t}}^{h_{t-12,t}}$ donné dans (10). Définissons les écarts-types

$$S_{h_{t-12,t}}^{h_{t-12,t}} = \sqrt{\frac{1}{n_{h_{t-12,t}}^{h_{t-12,t}}} \sum_{h_{t-12,t}} (O_{h_{t-12,t}}^{h_{t-12,t}} - \bar{O}_{h_{t-12,t}}^{h_{t-12,t}})^2} \quad (m = 0, 1, 2).$$

développement en série de Taylor de premier ordre d'un ratio de deux estimateurs, C'est-à-dire,

$$\begin{aligned} \text{var}(\hat{g}_{t-12}^{h\ell}) &= \text{var} \left\{ \hat{O}_{t-12}^{h\ell} \right\} \\ &\approx \frac{(\text{var}(\hat{O}_{t-12}^{h\ell}) - G_{t-12}^{h\ell} \hat{O}_{t-12}^{h\ell})}{(\text{var}(\hat{O}_{t-12}^{h\ell}) + (G_{t-12}^{h\ell} \hat{O}_{t-12}^{h\ell}) - 2G_{t-12}^{h\ell} \text{cov}(\hat{O}_{t-12}^{h\ell}, \hat{O}_{t-12}^{h\ell}))}. \end{aligned} \quad (2)$$

Le problème principal est l'estimation de $\text{cov}(\hat{O}_{t-12}^{h\ell}, \hat{O}_{t-12}^{h\ell})$. Aux sections suivantes, nous examinons ce terme et son estimation.

3.2 Terme de covariance du taux de croissance annuel

En utilisant le plan d'échantillonnage stratifié, nous pouvons écrire le terme $\text{cov}(\hat{O}_{t-12}^{h\ell}, \hat{O}_{t-12}^{h\ell})$ provenant de (2) sous la forme

$$\text{cov}(\hat{O}_{t-12}^{h\ell}, \hat{O}_{t-12}^{h\ell}) = \text{cov} \left(\sum_H \sum_{h=1}^H N_{t-12}^{h\ell} \text{cov}(\hat{O}_{t-12}^{h\ell}, \hat{O}_{t-12}^{h\ell}), \sum_H \sum_{h=1}^H N_{t-12}^{h\ell} \text{cov}(\hat{O}_{t-12}^{h\ell}, \hat{O}_{t-12}^{h\ell}) \right) \quad (3)$$

où $\hat{O}_{t-12}^{h\ell}$ désigne la moyenne d'échantillon du chiffre d'affaires dans la strate h pendant le mois $t - m$ ($m = 0, 12$). Notons que la stratification des unités au mois $t - 12$ peut différer de celle au mois t . Comme nous l'avons vu à la section 2.2, le renouvellement classique du panel a lieu en janvier. En outre, chaque établissement est affecté à la strate h correcte en fonction de son nombre réel d'employés en janvier ($h = 1, \dots, H$). Afin de tenir compte de ces caractéristiques du plan d'échantillonnage, définissons

$N_{t-12,t}^{h\ell}$: la taille de la sous-strate $U_{t-12,t}^{h\ell}$, c'est-à-dire l'ensemble d'unités qui, au mois $t - 12$, appartenait à la strate h et, au mois t , à la strate ℓ ($h, \ell = 1, \dots, H$) ;
 $O_{t-m}^{h\ell}$: le total de population de sous-strate du chiffre d'affaires dans $U_{t-12,t}^{h\ell}$ au mois $t - m$ ($m = 0, 12$) ;
 $\bar{O}_{t-m}^{h\ell}$: la moyenne de population de sous-strate du chiffre d'affaires dans $U_{t-12,t}^{h\ell}$ au mois $t - m$ [c'est-à-dire $\bar{O}_{t-m}^{h\ell} = O_{t-m}^{h\ell} / N_{t-12,t}^{h\ell}$, $m = 0, 12$] ;
 $n_{t-m}^{h\ell}$: la taille de l'échantillon $s_{t-12,t}^{h\ell}$ au mois $t - m$ ($0 \leq m \leq 12$) ;

En plus de la notation de la section 2, définissons la strate auxiliaire 0 pour les nouvelles unités aux mois $t - 12, \dots, t - 1$ et, de même, la strate $H + 1$ pour les unités disparues durant cette période. Nous pouvons alors écrire

$$\begin{aligned} \bar{O}_{t-12}^{h\ell} &= \sum_{H+1}^{H+1} \frac{n_{t-12}^{h\ell}}{n_{t-12}^{h\ell}} \bar{O}_{t-12}^{h\ell} \\ \bar{O}_{t-12}^{h\ell} &= \sum_{H+1}^{H+1} \frac{n_{t-12}^{h\ell}}{n_{t-12}^{h\ell}} \bar{O}_{t-12}^{h\ell} \end{aligned}$$

respectivement ($1 \leq h, \ell \leq H$). D'où, les covariances données par (3) peuvent être réécrites

$$\text{cov}(\bar{O}_{t-12}^{h\ell}, \bar{O}_{t-12}^{h\ell}) = \text{cov} \left(\sum_{H+1}^{H+1} \frac{n_{t-12}^{h\ell}}{n_{t-12}^{h\ell}} \bar{O}_{t-12}^{h\ell}, \sum_{H+1}^{H+1} \frac{n_{t-12}^{h\ell}}{n_{t-12}^{h\ell}} \bar{O}_{t-12}^{h\ell} \right) \quad (4a)$$

où nous avons utilisé $\text{cov}(n_{t-12}^{h\ell} \bar{O}_{t-12}^{h\ell}, n_{t-12}^{h\ell} \bar{O}_{t-12}^{h\ell}) = 0$ ($k \neq h$) et $\text{cov}(n_{t-12}^{h\ell} \bar{O}_{t-12}^{h\ell}, n_{t-12}^{h\ell} \bar{O}_{t-12}^{h\ell}) = 0$ ($g \neq \ell$). Cette dernière covariance est nulle parce que

$$\begin{aligned} \text{cov}(n_{t-12}^{h\ell} \bar{O}_{t-12}^{h\ell}, n_{t-12}^{h\ell} \bar{O}_{t-12}^{h\ell}) &= E \text{cov}(n_{t-12}^{h\ell} \bar{O}_{t-12}^{h\ell} | n_{t-12}^{h\ell}, n_{t-12}^{h\ell}) \\ &+ \text{cov}\{E(n_{t-12}^{h\ell} \bar{O}_{t-12}^{h\ell} | n_{t-12}^{h\ell}, n_{t-12}^{h\ell}), E(n_{t-12}^{h\ell} \bar{O}_{t-12}^{h\ell} | n_{t-12}^{h\ell}, n_{t-12}^{h\ell})\} \\ &= 0 + \bar{O}_{t-12}^{h\ell} \text{cov}(n_{t-12}^{h\ell}, n_{t-12}^{h\ell}) = 0. \end{aligned}$$

À la dernière ligne, nous avons également utilisé le fait que, pour $1 \leq g \leq H + 1$,

$$\text{cov}(n_{t-12}^{h\ell}, n_{t-12}^{h\ell}) = 0. \quad (5)$$

Pour une justification et les hypothèses qui sous-tendent (5), voir l'annexe A. En outre, à l'annexe A, nous proposons une autre méthode d'estimation quand cette variance n'est pas négligeable. La covariance dans (4b) peut s'exprimer

la disparition d'unités dans la population. Notons qu'en d'échantillonnage réelle f_h ne dépend pas du mois t . En fait, la procédure de mise à jour de s_t^h au mois t est la suivante. Soit $U_{t-1,t}^{0h}$ l'ensemble de nouvelles unités dans la strate h au mois $t-1$ et désignons sa taille par $N_{t-1,t}^{0h}$. Le nombre d'unités échantillonnées à partir de $U_{t-1,t}^{0h}$ au mois t est $n_{t-1,t}^{0h} = f_h N_{t-1,t}^{0h}$. En outre, désignons la différence requise $n_h^{0h} - n_{t-1,t}^{0h}$ par $n_{h,REQ}^{0h}$ et définissons $s_{t-1,t}^{h,PRE}$ par $s_{t-1,t}^{h,PRE} = s_{t-1,t}^{h,PRE} \cap U_{t-1,t}^{0h}$, c'est-à-dire l'ensemble de des unités dans l'échantillon $s_{t-1,t}^{h,PRE}$ à la période précédente qui sont encore présentes au mois t . Soit $n_{t-1,t}^{h,PRE}$ la taille de $s_{t-1,t}^{h,PRE}$. Quand $n_{h,PRE}^{0h} \geq n_{t-1,t}^{h,PRE}$ on abandonne aléatoirement un nombre d'unités correspondant à la différence; autrement, on sélectionne un nombre d'unités correspondant à la différence à partir de $U_{t-1,t}^{0h} \setminus s_{t-1,t}^{h,PRE}$. Notons que les unités supprimées de l'échantillon au mois $t-1$ ou avant cela peut être à nouveau sélectionnées au mois t .

2.2 Mise à jour annuelle

Chaque année en janvier, l'échantillon est mis à jour pour tenir compte à la fois de la restructuration des unités et du remplacement de 10 % de l'échantillon. Toutes les unités présentes dans l'échantillon de décembre qui existent encore en janvier sont stratifiées en fonction de leur taille réelle, c'est-à-dire le nombre d'emplois et le code de la CTI en janvier. Les limites des classes de taille proprement dites ne changent pas. Par conséquent, l'échantillon provenant d'une strate conformément à la nouvelle stratification de janvier peut contenir des unités ayant des probabilités d'inclusion différentes, parce que les unités migrent entre les strates ayant des fractions d'échantillonnage différentes.

Afin d'apporter une correction pour tenir compte des probabilités d'inclusion éventuellement différentes dans la strate h , désignons la sous-strate constituée des unités qui appartiennent à la strate h en décembre et à la strate ℓ en janvier par $U_{dec,janv}^{h\ell}$ et désignons sa taille par $N_{dec,janv}^{h\ell}$ ($h, \ell = 1, \dots, H$). Par analogie avec la procédure de mise à jour mensuelle, définissons $s_{dec,janv}^{h\ell,PRE}$ par $s_{dec,janv}^{h\ell,PRE} = s_{dec,janv}^{h\ell,PRE}$ et désignons par $n_{dec,janv}^{h\ell,PRE}$ la taille de $s_{dec,janv}^{h\ell,PRE}$. Puisque la taille requise de l'échantillon $s_{dec,janv}^{h\ell,REQ}$ provenant de $U_{dec,janv}^{h\ell}$ en janvier est $f_{dec,janv}^{h\ell} = f_{dec,janv}^{h\ell} N_{dec,janv}^{h\ell}$, la mise à jour annuelle de l'échantillon $s_{dec,janv}^{h\ell,PRE}$ s'effectue comme il suit.

Premièrement, quand $n_{dec,janv}^{h\ell,PRE} \geq n_{dec,janv}^{h\ell,REQ}$, un nombre d'unités correspondant à la différence est supprimé aléatoirement de $s_{dec,janv}^{h\ell,PRE}$. En outre, 10 % des $n_{dec,janv}^{h\ell,REQ}$ unités encore présentes dans $s_{dec,janv}^{h\ell,PRE}$ sont remplacées par des unités provenant de $U_{dec,janv}^{h\ell,PRE} \setminus s_{dec,janv}^{h\ell,PRE}$ à condition que ce dernier ensemble contienne suffisamment d'unités. Si le nombre d'unités disponibles n'est pas suffisant, le nombre d'unités remplacées est seulement $N_{dec,janv}^{h\ell,PRE} - n_{dec,janv}^{h\ell,PRE}$. Deuxièmement, quand

3. Variance du taux de croissance annuel du chiffre d'affaires mensuel

3.1 Variance du taux de croissance annuel

Soit O_t^i le chiffre d'affaires total de tous les établissements comptés dans la population au mois t et $g_{t,s}^{i,s}$ la variation relative du niveau du chiffre d'affaires entre les mois t et s , c'est-à-dire

$$g_{t,s}^{i,s} = \frac{O_t^i}{O_s^i} - 1 \quad (t > s).$$

Pour les estimations correspondantes, il est vérifié par définition que

$$g_{t,t}^{i,s} = \frac{O_t^i}{O_t^i} - 1, \quad (1)$$

où un « chapeau » indique qu'il s'agit d'une estimation; pour un estimateur, nous utilisons la même notation. En outre, définissons

$$G_{t,t-12}^{i,s} \equiv \frac{O_t^{i-12}}{O_t^i} = 1 + g_{t,t-12}^{i,s}.$$

Afin d'estimer la variance du taux de croissance annuel du chiffre d'affaires mensuel, nous utilisons le

Toutes les unités statistiques ou *etablisements* figurent sur la liste du Register général des entreprises (RGE) tenu à jour par Statistics Netherlands. Le registre est mis à jour chaque mois au moyen de sources administratives en ce qui concerne les créations et les disparitions d'établissements, et une fois par an, le 31 décembre, en ce qui concerne la catégorie de taille et le type d'activité économique (code de la CTE). Il convient de souligner que l'enregistrement dans le RGE peut avoir lieu avec un certain retard par rapport aux changements dans la population (création d'établissements, disparition d'établissements, changements de classe de taille, etc.). De surcroît, les unités disparues (inconnues) dans la base de sondage peuvent entraîner un biais dans les estimations du niveau du chiffre d'affaires. Afin d'éviter ce genre de biais, il est important de détecter rapidement les disparitions et de supprimer ces unités de la base de sondage. Les disparitions d'établissements décelées dans l'échantillon peuvent jouer un rôle ici. Cependant, une analyse plus approfondie et la correction de ces erreurs dépassent le cadre du présent article sur l'estimation de la variance des taux de croissance. Pour estimer ces variances, nous supposons que les unités de la population et leurs caractéristiques dans le registre sont correctes. De même, nous supposons que la non-réponse aux enquêtes est nulle.

Chaque premier jour du mois, un échantillon EASS est tiré du RGE pour estimer le chiffre d'affaires du mois courant. En fait, on utilise un échantillon rotatif. L'échantillon est stratifié selon la taille et selon le type d'activité économique. La probabilité réelle de sélection dépend de la taille et de l'activité économique. La probabilité de sélection augmentée avec la taille de l'établissement, les plus grands étant inclus dans l'échantillon avec une probabilité de 1. Pour certains codes de la CTE, il existe non seulement des données d'enquête, mais aussi des données provenant de sources administratives. Les unités déjà présentes dans les fichiers administratifs sont considérées comme une strate distincte. Les estimations pour cette strate ont une variance nulle.

L'échantillon est mis à jour de deux façons. Chaque mois, il est corrigé pour tenir compte des nouvelles unités et des unités disparues dans la population. Une fois par an, en janvier, 10 % des unités de l'échantillon sont remplacées et des corrections sont apportées aux strates. Nous allons maintenant discuter plus en détail des mises à jour mensuelles et annuelles.

2.1 Mise à jour mensuelle

Chaque mois t ($t = 1, 2, \dots$), une proportion fixe f_h des N_h^t unités de la strate U_h^t sont échantillonnées ($h = 1, \dots, H$). Cela donne un échantillon s_h^t de taille $n_h^t = f_h N_h^t$. Donc, le nombre réel d'unités dans l'échantillon peut changer d'un mois à l'autre à cause de la création et de

Chaque mois, Statistics Netherlands estime le chiffre d'affaires mensuel pour certains des codes principaux de la CTE. Les données publiées comprennent les taux de croissance sur 12 mois du chiffre d'affaires mensuel, c'est-à-dire la variation relative du niveau mensuel du chiffre d'affaires comparativement à 12 mois plus tôt. Dans la suite de l'exposé, nous parlerons de ce taux de croissance comme étant le taux de croissance annuel.

2. Le plan d'échantillonnage des enquêtes-entreprises des Pays-Bas

La présentation de l'article est la suivante. À la section 2, nous décrivons brièvement l'enquête-entreprise sur le plan d'échantillonnage. À la section 3, nous établissons les formules de la variance de l'estimateur d'une variation. À la section 4, nous illustrons la procédure d'estimation de la variance en comparant les variances de deux estimateurs différents du taux de croissance annuel du chiffre d'affaires mensuel des supermarchés des Pays-Bas au cours de la période 2003-2004. Enfin, à la section 5, nous résumons les principaux résultats et présentons nos conclusions.

Pays-Bas pour des périodes de quatre semaines. La variance annuelle du chiffre d'affaires des supermarchés des Pays-Bas pour des périodes de quatre semaines. Afin de clarifier la procédure d'estimation de la variance, nous décrivons son application au taux de croissance annuel du chiffre d'affaires des supermarchés des Pays-Bas pour des périodes de quatre semaines.

Dans le présent article, nous établissons les expressions requiert certaines opérations algébriques matricielles. conditionnellement à la taille d'échantillon par strate, ce qui formules en se basant sur un échantillonnage de Poisson périodes différentes est fixe, Berger (2004) établit ses taille du chevauchement des deux échantillons à deux requiert certains calculs algébriques. En supposant que la des formules en utilisant des indicateurs d'inclusion qui plan d'échantillonnage de la Suède, Nordberg (2000) établit comprenant des unités qui migrent entre les strates. Pour le situation plus compliquée d'une population dynamique (2000) et Berger (2004) établissent des formules pour la fication, mais sans modification des strates. Nordberg (1995) considèrent des populations dynamiques et la strat-compte de la stratification. Hidroglou, Särndal et Binder traitent de l'estimation de la variance d'une variation dans des populations dynamiques, mais ils ne tiennent pas

À propos de la variance des variations estimées d'après des panels rotatifs et des strates dynamiques

Paul Knothnerus et Arnout van Delden¹

Résumé

De nombreuses enquêtes-entreprises fournissent des estimations du chiffre d'affaires mensuel pour les principaux codes de la Classification type des industries. Cela inclut les estimations des variations du niveau du chiffre d'affaires mensuel comparativement à 12 mois plus tôt. Comme des échantillons chevauchants sont souvent utilisés dans les enquêtes-entreprises, les estimations du chiffre d'affaires durant des mois consécutifs sont corrélées, ce qui complique le calcul de la variance des variations. Le présent article décrit une procédure générale d'estimation de la variance qui comprend des corrections annuelles des strates quand des échantillonnages passent dans d'autres strates en raison de leur taille réelle. La procédure tient également compte du renouvellement des échantillons, ainsi que des nouvelles unités et des unités disparues. L'article se termine par un exemple de calcul de la variance de l'estimation du taux de croissance annuel du chiffre d'affaires mensuel des supermarchés des Pays-Bas.

Mots clés : Covariances conditionnelles ; corrections des strates ; échantillons chevauchants ; enquêtes-entreprises ; nouvelles unités ; unités disparues.

1. Introduction

Dans de nombreuses enquêtes, une population en évolution est échantillonnée de manière répétée afin de pouvoir estimer le niveau et la variation de niveau d'une caractéristique entre deux éditions de l'enquête. Par exemple, beaucoup de pays réalisent une enquête-entreprise mensuelle pour estimer le niveau du chiffre d'affaires mensuel et la variation de ce niveau comparativement à un mois ou à une année plus tôt ; voir Konoschnik, Monsoor et Dettelsen (1985). Un autre exemple est celui de l'enquête sur la population active qui comprend l'échantillonnage mensuel de la population pour estimer le nombre de personnes occupées et le taux de chômage. L'estimation de la variance est nécessaire afin de savoir si les variations observées sont statistiquement significatives. L'estimation de la variance est également nécessaire à l'étape de la conception de l'enquête, afin de déterminer la taille et la répartition optimales de l'échantillon, ou l'estimateur optimal.

Dans les enquêtes répétées, on estime souvent les changements s'appuyant sur une stratification de la population. Les entreprises sont très hétérogènes en ce qui concerne la taille et le type d'activité économique. Par conséquent, les enquêtes-entreprises sont habituellement réalisées selon un plan d'échantillonnage aléatoire simple stratifié (FAS) avec tirage sans remise ; voir Smith, Pont et Jones (2003). Dans les enquêtes auprès des ménages ou des particuliers, l'échantillon n'est en général pas stratifié, parce que les ménages sont moins hétérogènes. Certaines enquêtes sociales, telles que les enquêtes sur la population active, comprennent toutefois une poststratification afin de réduire la variance et le biais de l'estimateur.

En vue d'établir les formules de la variance d'une variation estimée dans une population comprenant des strates dynamiques, il convient de faire attention à trois facteurs de complication. Premièrement, la variation de niveau est le résultat de deux composantes. L'une est due à la variation de la moyenne de population des unités qui se trouvaient dans la même strate lors des deux éditions de l'enquête. L'autre est causée par le changement de composition de la strate entre deux éditions de l'enquête résultant de la création de nouvelles unités et de la disparition d'unités existantes au sein de la population, ainsi que de la migration des unités de population entre les strates ; voir Holt et Skinner (1989). Deuxièmement, en raison de la migration de certaines unités de population entre les strates, la moyenne estimée de la strate h à la période t peut être corrélée avec la moyenne de la strate h à la période $t + 1$. Troisièmement, une autre complication tient au fait que la population est échantillonnée répétitivement, ce qui donne des échantillons partiellement chevauchants entre deux périodes. Différents plans à panel rotatif peuvent être utilisés dans les enquêtes-entreprises.

Divers auteurs ont établi des formules d'estimateurs de la variance fondées sur le plan de sondage pour l'estimation des variations. En supposant une grande population sans création ni disparition d'unités, Kish (1965) a dérivé une expression pour la variance d'une variation estimée fondée sur des échantillons chevauchants. Tam (1984) a éliminé l'hypothèse d'une grande population. Partant des résultats de Tam, Qualité et Tillé (2008) comparent plusieurs estimateurs de variance d'une variation estimée. Wood (2008) généralise les résultats de Tam aux enquêtes avec probabilités de sélection inégales. Lowrer (1979) et Laniet (1987)

couvert est supposé être un échantillon de Poisson tiré de l'univers souhaité, l'échantillon est un échantillon aléatoire simple tiré de l'univers couvert, et les répondants représentent un échantillon de Poisson tiré de l'échantillon original.

Bibliographie

Banker, M.D., Rathwell, S. et Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 canadian census. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 764-769.

Chang, T., et Kott, S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 557-571.

Clark, R., et Chambers, R.L. (2008). Adaptive Calibration for Prediction of Finite Population Totals. University of Wollongong (document de travail en ligne).

Falk, G. (2010). *Calibration Adjustment for Nonresponse in Cross-Classified Data*, University of Virginia (dissertation).

Fuller, W.A., Loughin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la *Nationwide Food Consumption Survey* de 1987-1988. *Techniques d'enquête*, 20, 79-89.

Kott, S., et Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105, 1265-1275.

Sämdal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Silva, P.L.D.N., et Skinner, C.J. (1997). Sélection des variables pour l'estimation par régression dans le cas des populations finies. *Techniques d'enquête*, 23, 25-35.

nous fixons la valeur de b correspondant à zéro. Dans les deux derniers cas, nous désignons les entrées de \mathbf{B} par b_{ij} . Pour $k \in S_{ij}$, soit $u_k = w_{ij}(y_k - b_{ij})$ pour les estimateurs à marges calées et par ratissage, et $u_k = w_{ij}(y_k - b_{ij})$ pour les estimateurs à fréquences de cellule calées et à fréquences de cellule calées exponentiel.

Essentiellement, Chang et Kott (2008) ont montré qu'asymptotiquement, l'estimateur calé a la même forme qu'un estimateur par la régression de la forme correspondant à l'équation (6.6.1) de Sämdal et coll. (1992), où la matrice \mathbf{B} susmentionnée joue le rôle du \mathbf{B} dans (6.6.1) et les poids d'échantillonnage d_k sont remplacés par $d_k f(\mathbf{z}_k^T \beta)$. Pour l'échantillonnage sans remise, ils proposent d'estimer la variance de \hat{t}_{xyfV} en apportant les changements analogues à l'équation (6.6.3) de Sämdal et coll. (1992).

Pour l'échantillonnage aléatoire simple et en l'absence de non-réponse et de non-couverture, l'estimateur de variance devient

$$\hat{V} = \frac{n}{N^2} (1 - n/N) s_z^2 \tag{15}$$

où s_z^2 est la variance d'échantillon des u_k . En présence de non-réponse, si l'on suppose que les répondants S représentent un échantillon de Poisson tiré de l'échantillon aléatoire simple original avec les probabilités poissonniennes $f(\mathbf{z}_k^T \beta_0)^{-1}$, l'estimateur de variance devient

$$\hat{V} = \frac{n}{N^2} (1 - n/N) s_z^2 + \frac{n}{N} \sum_i \sum_j (1 - w_{ij}) \sum_{k \in S_{ij}} u_k^2 \tag{16}$$

où s_z^2 est la variance d'échantillon des u_k . La même formule fonctionne pour la non-couverture où $f(\mathbf{z}_k^T \beta_0)^{-1}$ représente la probabilité de couverture et de réponse combinées dans un modèle à trois degrés dans lequel l'univers

Dans nos approches de calage, nous supposons que ces probabilités sont une fonction des classifications par ligne et par colonne. Quand des cellules sont regroupées sans maintenir la classification double, ces probabilités sont difficiles à interpréter.

Remerciements

L'auteur remercie Phil Kott et John Eltinge de lui avoir fourni plusieurs éclaircissements très intéressants. Il remercie aussi Larry Huff et Ken Robertson du Bureau of Labor Statistics de l'avoir aidé à obtenir et à comprendre les données.

Annexe

Ici, nous établissons, en utilisant les équations (16) et (17) de Chang et Kott (2008), les estimateurs de variance fondés sur l'échantillon pour les quatre estimateurs étudiés à la section 2.

Soit

$$\hat{H}_y = \frac{\partial f_{yzjv}}{\partial \beta}(\beta).$$

Ici f_{yzjv} est défini en (2). H_y est un vecteur ligne contenant une entrée pour chaque variable Z . Dans le cas de la suppression d'une unité aberrante, H_y contient $(I + J - 1)$ entrées, une pour chacune des $I + J - 1$ variables indicatrices linéairement indépendantes pour les classifications par ligne et par colonne.

Pour les estimateurs à marges calées et à fréquences de cellule calées, $f(\eta) = 1 + \eta$. Définissons les constantes s_{ij}

et t_{ij} par

$$s_{ij} = \frac{u}{N} r_{ij} \quad t_{ij} = \frac{u}{N} \sum_{k \in s_{ij}} \frac{u}{y_k}$$

Alors, un simple calcul montre que, s'il existe une entrée dans H_y pour la i^e classification par ligne, nous plaçons $\sum_j t_{ij}$ dans cette entrée. De même, si une entrée existe pour la j^e classification par colonne, nous plaçons $\sum_i t_{ij}$ dans cette entrée. Ici, nous utilisons la convention voulant que, si la i^e ligne ou la j^e colonne n'est pas l'une des $I + J - 1$ variables indicatrices linéairement indépendantes choisies, le β_i ou β_j correspondant est égal à 0.

Pour l'estimateur par ratisage (*raking*) et l'estimateur à fréquences de cellule calées exponentiel, $f(\eta) = e^\eta$ et nous pouvons calculer H_y de manière similaire en utilisant

$$s_{ij} = \frac{u}{N} \exp(\beta_i + \beta_j) r_{ij} \quad t_{ij} = \frac{u}{N} \sum_{k \in s_{ij}} \exp(\beta_i + \beta_j) y_k$$

Ici, nous adoptons la convention voulant que, si la i^e ligne ou la j^e colonne n'est pas l'une des $I + J - 1$ variables indicatrices linéairement indépendantes choisies, le β_i ou β_j correspondant est égal à 1.

Par analogie avec (2), soit

$$f_{kxzjv} = \sum_{k \in S} d_k f(Z_k^T \beta) x_k.$$

f_{kxzjv} est un vecteur colonne contenant une entrée pour chaque variable x . Définissons la matrice H comme étant

$$\hat{H} = \frac{\partial f_{kxzjv}}{\partial \beta}(\beta).$$

H est une matrice contenant une ligne pour chaque variable x et une colonne pour chaque variable z .

Pour les estimateurs à fréquences de cellule calées et à fréquences de cellule calées exponentiel, la matrice H est de dimensions $IJ \times (I + J - 1)$. Chaque des lignes de H correspond à une paire (i, j) de classifications par ligne et colonne. Nous plaçons s_{ij} dans la ligne correspondant à (i, j) et dans les colonnes correspondant à la i^e classification par ligne et à la j^e classification par colonne (quand ces colonnes existent). Toutes les autres entrées de H sont

fixées à zéro.

Pour les estimateurs à marges calées et par ratisage, la matrice H est de dimensions $(I + J - 1) \times (I + J - 1)$. Si une ligne (et donc une colonne) de H existe pour la i^e classification par ligne, nous plaçons $\sum_j s_{ij}$ dans l'entrée diagonale correspondant de H . De même, si une ligne et une colonne existe pour la j^e classification par colonne, nous plaçons $\sum_i s_{ij}$ sur la diagonale de H . Nous plaçons s_{ij} dans l'entrée dont la ligne correspond à la i^e classification par ligne et dont la colonne correspond à la j^e classification par colonne (si celles-ci existent). Nous plaçons aussi s_{ij} dans l'entrée dont la colonne correspond à la i^e classification par ligne et dont la ligne correspond à la j^e classification par colonne (de nouveau si celles-ci existent). Les valeurs de toutes les autres entrées de H sont fixées à zéro.

Soit $B = H^T (H^T V^{-1} H)^{-1} H^T V^{-1}$ où, pour le moment, nous utilisons une matrice identité pour V . B est de dimensions $(I + J - 1) \times (I + J - 1)$ pour les estimateurs à marges calées et par ratisage, et de dimensions $I \times IJ$ pour les estimateurs à fréquences de cellule calées exponentiel. Dans les deux premiers cas, nous désignons les entrées de B par b_i ou b_{ij} , et pour le cas unitaire où un indice de colonne ou de ligne ne correspond pas à une des $I + J - 1$ variables indicatrices indépendantes, l'une des $I + J - 1$ variables indicatrices indépendantes, l'une des

Tableau 7
Comparaison empirique de quatre estimateurs

Estimateur	Biais	Erreur-type	REQM	Biais	Erreur-type	REQM
n = 200						
Non pondéré	644	1 006 956	1 006 956	-9 970	1 481 450	1 481 483
Poststratifié	-5 387	1 026 266	1 026 280	-2 149	1 548 833	1 548 834
Fréq. cell. calées	-224 198	942 164	968 472	-203 531	1 377 823	1 392 775
Fréq. cell. cal. pondéré	-248 937	919 419	952 523	-232 558	1 326 234	1 346 469
n = 1 000						
Non pondéré	-3 317	445 676	448 228	1 544	679 148	679 150
Poststratifié	-2 967	448 218	445 687	1 672	685 370	685 372
Fréq. cell. calées	-54 311	436 821	440 185	-44 942	665 799	667 314
Fréq. cell. cal. pondéré	-63 327	432 396	437 008	-54 913	660 726	663 004
n = 5 000						
Non pondéré	2 466	206 249	206 264	-2 539	304 852	304 863
Poststratifié	2 108	205 661	205 672	-2 705	304 751	304 763
Fréq. cell. calées	-8 265	204 693	204 859	-12 096	303 231	304 472
Fréq. cell. cal. pondéré	-10 551	204 080	204 352	-14 697	302 311	302 668

6. Conclusion

L'utilisation dans l'expression (3) de matrices de pondération $V(\beta)^{-1}$ qui dépendent de β n'a pas été examinée dans le présent article. L'expérimentation avec ce genre de matrices n'a pas été encourageante. Le temps de calcul augmentait considérablement et, dans un nombre important de cas, la convergence numérique n'était pas atteinte, sans aucune amélioration de l'efficacité par rapport aux estimateurs avec matrice V fixe considérée ici. L'auteur n'a peut-être pas choisi la bonne matrice $V(\beta)$, pour ses essais.

Outre la fonction de lien rétrotransformée exponentielle, l'auteur a essayé la fonction de lien rétrotransformée logistique $f(\eta) = (1 + e^{-\eta})^{-1}$. Ces exécutions n'ont pas convergé non plus. Tout bien réfléchi, la raison en est évidente : comme il n'y avait aucune non-réponse ni aucun problème de non-couverture dans les simulations, les ajustements des poids de calage $f(z^T\beta) \rightarrow 1$ quand $n \rightarrow \infty$. Cependant, l'n'est pas l'image de f . Il convient de souligner que, dans Chang et Kott (2008), une fonction de lien rétrotransformée logistique a été utilisée pour corriger la non-réponse.

Plusieurs questions évidentes se posent. Par exemple, comment les résultats de la présente étude évolueraient-ils si l'on utilisait un plan de sondage plus compliqué que l'échantillonnage aléatoire simple, ou s'il se produisait une non-réponse et/ou une non-couverture et que le calage était utilisé pour la corriger ? Falk (2010) examine ces questions tant théoriquement qu'en exécutant d'autres simulations au moyen de la population du QCEW. Falk (2010) considère également des fonctions de lien non linéaires.

Les extensions à des données croisées à trois dimensions (et plus) sont évidentes. Si I, J, K désignent le nombre de

cellules dans chacune des trois classifications, il existe IJK cellules entièrement classées dont les totaux peuvent être utilisés comme variables x de référence. Il existe $IJ + IK + JK - I - J - K + 1$ variables marginales à une et à deux dimensions qui peuvent être utilisées pour les variables z du modèle. Il est clair que l'on pourrait ne pas vouloir utiliser la pléthore de variables disponibles.

Dans le contexte du calage linéaire en utilisant les mêmes variables x et z , plusieurs études ont porté sur le choix des variables. Les études de Banker, Rathwell et Majkowski (1992), Silva et Skinner (1997), et de Clark et Chambers (2008) en sont des exemples. Les auteurs du dernier article font remarquer qu'un trop grand nombre de variables peut détériorer l'EQM de T_p .

Les options de remplacement de la poststratification examinées ici peuvent être utilisées en présence de cellules vides. Par exemple, dans nos simulations, la fréquence prévue dans la cellule définie par l'état E et le groupe d'industries 5, est de 0,36 quand $n = 200$. On pourrait être tenté de regrouper certaines cellules et d'utiliser la poststratification. Toutefois, en général, il n'est pas possible de le faire et de maintenir la structure à deux dimensions des données. Comme la poststratification, notre approche comprend l'introduction de poids pour équilibrer l'échantillon, mais évite le regroupement de cellules. Ces approches augmentent généralement le biais, mais peuvent donner lieu à une réduction importante de l'EQM.

En outre, en présence de non-réponse ou de non-couverture, les inverses des ajustements des pondérations peuvent être considérés, sous un modèle de quasi-randomisation pour la réponse ou la couverture, comme probabilités estimées de réponse et/ou de couverture.

Nous donnons à l'estimateur \hat{f}_{xy_0} de l'équation (14) le nom d'estimateur à fréquences de cellule calées pondéré.

Des simulations exécutées en se servant de variables réponses artificielles y , bien que non présentes, montrent que, si le modèle (13) est vérifié, l'estimateur à fréquences de cellule calées pondéré \hat{f}_{xy_0} donne de nettement meilleurs résultats que les autres estimateurs pris en considération ici. Le tableau 6 donne les statistiques pour l'estimateur \hat{f}_{xy_0} pour les populations et les variables étudiées aux tableaux 2 à 5.

Comparativement aux tableaux 2 à 5, nous voyons que, dans tous les cas, \hat{f}_{xy_0} est, de tous les estimateurs considérés, celui qui possède le biais le plus élevé, mais l'EQM la plus faible. Pour $n = 200$ et la population complète, \hat{f}_{xy_0} produit un gain d'efficacité (mesuré par l'EQM) de 14,8 %, comparativement aux résultats empiriques obtenus pour l'estimateur sans biais pour l'estimation de l'emploi total et un gain d'efficacité de 21,1 % pour l'estimation de la rémunération totale. Pour $n = 200$ et la population dont a été supprimée une unité aberrante unique, les gains d'efficacité correspondants sont de 14,2 % et de 21,7 % pour l'emploi total et la rémunération totale, respectivement.

Le rédacteur a proposé que nous comparions nos estimateurs regroupées pour éviter le problème de cellules vides dans l'échantillon. Nous avons examiné cette question pour une taille d'échantillon $n = 200$ pour laquelle il est fort probable qu'il existe des cellules vides. Nous avons créé 14 poststrates. Neuf de ces poststrates correspondaient aux neuf plus grandes cellules dans les données originales. Les cinq autres poststrates étaient A1 et A2 ; A3, A5 et B4 ; A4, B5 et C4 ; C5 et D4 ; et toutes les cellules provenant de l'État E regroupées avec D5. Après ces combinaisons, les cinq poststrates combinées avaient une taille variant de 4,07 % à 5,06 % de la population, et les neuf cellules originales retenues avaient une taille variant de 4,62 % à 11,47 % de la population.

Malheureusement, l'auteur n'a plus accès à la base de données du QCEW. À part les fréquences de cellule du tableau 1, il ne dispose que des moyennes, des écarts-types et des valeurs maximales par cellule. Il a construit une pseudo population en utilisant les carrés de variables gamma générées aléatoirement. Les variables gamma au carré ont été créées de manière à ce qu'elles aient les mêmes moyennes et écarts-types de cellule que les moyennes et écarts-types de cellule des données originales. Ensuite, les variables gamma au carré ont été arrondies vers le haut à une valeur entière. Pour ces pseudo populations,

Tableau 6
Statistiques empiriques pour \hat{f}_{xy_0} de l'équation (14)

n	Biais	Erreur-type	REQM	R.C. var. est.
200	-244 749	967 066	997 556	923 492
1 000	-64 839	490 758	495 023	483 550
5 000	-10 767	221 702	221 964	219 408
Population complète – Emploi total				
200	-242 528	1 388 489	1 409 511	1 333 793
1 000	-62 091	752 603	755 160	744 315
5 000	-9 821	332 682	332 827	333 782
Population avec unité aberrante supprimée – Emploi total				
200	-236 812	881 844	913 088	842 191
1 000	-67 468	405 215	410 793	396 105
5 000	-11 482	189 501	189 848	185 483
Population avec unité aberrante supprimée – Rémunération totale				
200	-228 441	1 194 922	1 216 562	1 151 417
1 000	-66 765	565 008	568 939	557 676
5 000	-11 138	263 699	263 934	260 768

De toute évidence, la poststratification n'est pas utile. Même si aucune poststrate n'avait une fréquence prévue inférieure à huit, la taille des poststrates réelles était assez variable. En outre, les populations de cellule sont relativement asymétriques de sorte que les moyennes d'échantillon des poststrates sont assez variables.

Les autres conclusions concernant les pseudo populations reflètent celles tirées pour les populations réelles. En particulier, pour $n = 200$ et la pseudo population pour l'emploi, l'estimateur à fréquences de cellule calées pondéré \hat{f}_{xy_0} donne un gain d'efficacité de 11,8 %, comparativement à l'estimateur sans biais. Pour la pseudo population pour la rémunération et $n = 200$, le gain d'efficacité est de 21,1 %.

$T_y = 3\,149\,491$ pour l'emploi et $2\,305\,273$, en dizaines de milliers de dollars, pour la rémunération.

Une distribution gamma au carré a été utilisée parce que la distribution gamma n'est pas suffisamment asymétrique avec étallement vers la droite, même si, dans presque toutes les cellules, la plus grande valeur dans la population originale est supérieure à la plus grande valeur dans la pseudo population. Naturellement, sans les données originales, nous ne pouvons pas faire la distinction entre un étallement de la distribution vers la droite et une tendance à produire des valeurs aberrantes.

En tout, 10 000 échantillons Monte Carlo ont été tirés pour chaque taille d'échantillon. Les résultats sont présentés au tableau 7. Pour l'estimateur poststratifié, cinq des échantillons de taille 200 possédaient une poststrate vide et ces exécutions ont été exclues des résultats du tableau 7.

Chang : Options de calage pour remplacement de données croisées à deux dimensions

$$(21) \quad {}^f\mathfrak{g} + {}^t\mathfrak{x} = \frac{{}^f f}{{}^f\theta}$$

En nous souvenant que θ_{ij} représente la valeur moyenne de la variable d'intérêt y dans la $(i, j)^{\text{e}}$ cellule, l'expression (12) ne semble pas être une approximation très prometteuse de la vérité. Une approximation plus probable serait le modèle d'ANOVA à deux critères de classification habituel

$$(\xi I) \quad \beta_f + \alpha = \theta_f$$

Supposons que nous remplacions les variables $\mathbf{x} = \mathbf{C}\mathbf{x}$ par certaines matrices diagonales \mathbf{C} . Notons que les lignes et les colonnes de \mathbf{C} sont doublement indexées par (i, j) et désignons par c_{ij} l'entrée diagonale dans les (i, j) ligne et colonne. Soit $\theta = \mathbf{C}^{-1}\theta_i$, de sorte que le modèle (6) peut être réécrit sous la forme

$$\cdot^k \mathbf{z} + \theta_{\mathbf{z}}^k \mathbf{x} = \cdot^k \mathcal{Y}$$

Maintenant, la matrice μ^{ex} contient C_{ij}^{eff} dans la $(i, j)^{\text{e}}$ ligne et dans les colonnes correspondant à i et à j . Ainsi, θ sera dans l'étendue des colonnes de μ^{ex} si et uniquement si

$$(\mathfrak{g} + \mathfrak{a})f\mathfrak{c} = \mathfrak{g} = \mathfrak{c}\mathfrak{g}.$$

Donc, le modèle (13) est équivalent à $c_{ij} = f_{ij}^{-1/2}$. Il est facile de vérifier que

$$\begin{aligned} \theta_{\perp}(\Lambda^{i_{\text{ZK}}^{\text{tr}}} \mathbf{d} - \mathbf{I})^{\text{x}} \mathfrak{Z}(\Lambda^{i_{\text{ZK}}^{\text{tr}}} \mathbf{d} - \mathbf{I})_{\perp} \theta \\ = \theta_{\perp}(\Lambda^{i_{\text{ZK}}^{\text{tr}}} \mathbf{d} - \mathbf{I})^{\text{x}} \mathfrak{Z}(\Lambda^{i_{\text{ZK}}^{\text{tr}}} \mathbf{d} - \mathbf{I})_{\perp} \theta \end{aligned}$$

quand $\mathbf{V} = \mathbf{C}^{-2}$. Nous proposons donc d'utiliser la matrice diagonale \mathbf{V}_o dont les entrées diagonales sont $f_o^{(j)}$. Avec ce choix de \mathbf{V}_o , l'équation (9) suggère l'estimateur pour l'échantillonnage aléatoire simple

$$(\mathfrak{t} \mathfrak{l}) \left(\mathfrak{y} \mathfrak{A}^{\mathfrak{y}} \mathfrak{z} \sum \frac{\mathfrak{u}}{\mathfrak{l}} \right)_{\mathfrak{l}} (\mathfrak{z} \mathfrak{y} \mathfrak{y} \mathfrak{l} \mathfrak{o} \mathfrak{A} \mathfrak{z} \mathfrak{y} \mathfrak{l}) \mathfrak{z} \mathfrak{y} \mathfrak{l} \mathfrak{o} \mathfrak{A} \mathfrak{z} (\mathfrak{s} \mathfrak{z} \mathfrak{N} - \mathfrak{x} \mathfrak{z}) + \mathfrak{s} \mathfrak{z} \mathfrak{N} = \mathfrak{o} \mathfrak{A} \mathfrak{z} \mathfrak{y} \mathfrak{l}$$

où $\hat{\mu}^x = n^{-1} \sum_{k \in S} x_k^T z_k^T$. Dans notre cas, μ^x ainsi que μ^y sont connues à partir de N_B^j , mais dans l'esprit de l'estimation par le ratio, il est préférable d'utiliser $\hat{\mu}^x$ au lieu de μ^x . Cette observation heuristique a été démontrée au moyen de simulations (non présentes) en se servant de la population du QCEW.

où Σ_x est la matrice de covariance de \mathbf{x} et

$$(01) \quad \cdot \mathbf{I} \mathbf{A}^{\mathbf{Z}} \mathbf{I} (\mathbf{Z} \mathbf{I} \mathbf{A}^{\mathbf{Z}} \mathbf{I})^{\mathbf{Z}} \mathbf{I} = \mathbf{A}^{\mathbf{Z}} \mathbf{I} \mathbf{d}$$

Maintenant,

$$\begin{aligned} & (\cdot|_{-}u)O + {}^J A {}^J f \sum \frac{u}{1} = \\ & u)O + [({}^s \underline{x} | {}^s \underline{f}) \Lambda \mathbb{V} \Lambda] \mathbb{E} = [({}^s \underline{x} | \Lambda^{2s} \mathbb{f}) \Lambda \mathbb{V} \Lambda] \mathbb{E} \\ & (\cdot|_{-}u)O + ({}^f \underline{x}) A \sum \frac{u}{1} = \\ & (\cdot|_{-}u)O + ({}^s \underline{x} | {}^s \underline{f}) \Lambda \mathbb{V} \Lambda = ({}^s \underline{x} | \Lambda^{2s} \mathbb{f}) \Lambda \mathbb{V} \Lambda \end{aligned}$$

Il est facile de voir que

$$\frac{1}{n} \theta^T \Sigma_x \theta = \text{Var}[E(\underline{y}_s | \underline{x}_s)] =$$

Puisque $\text{Var}[\hat{\mu}^{Y^s}] = \text{Var}[E[\hat{\mu}^{Y^s} | \underline{x}^s]] + E[\text{Var}[\hat{\mu}^{Y^s} | \underline{x}^s]]$ et de même pour $\text{Var}[\hat{f}^s]$, $\text{Var}[\hat{\mu}^{Y^s}]$, $\text{Var}[\hat{f}^s]$ et $\text{Var}[\hat{\mu}^{Y^s}]$ sont des termes d'ordre $O(n^{-1})$ quand

$$(11) \quad \theta^x \mathfrak{Z}_I \theta > \theta_{\mathcal{I}}(\Lambda^{x, \mathfrak{Z}_I} \mathbf{P} - \mathbf{I})^x \mathfrak{Z}(\Lambda^{x, \mathfrak{Z}_I} \mathbf{P} - \mathbf{I})_{\mathcal{I}} \theta$$

Les calculs établissent également que la contribution du biais quadratique à l'erreur quadratique moyenne de $\hat{\lambda}^{\text{AZ}}$ est asymptotiquement négligeable.

5. Une nouvelle matrice de pondération V^{-1} proposée

A la présente section, nous retournons aux variables de référence \mathbf{x} et du modèle \mathbf{z} \mathbf{z} originaux. Quand $\mathbf{V} = \mathbf{I}$, la matrice identité, $\mathbf{d}^{x,1}$ est la projection de θ sur l'élément que des colonnes de $\mathbf{H}^{x,1}$. Le premier membre de (11) sera nul si θ est dans cette étendue des colonnes.

Pour simplifier, nous écrivons $\mu^{\mathbf{xz}}$ sous forme d'une matrice singulière, de rang $I + J - 1$, contenant une ligne pour chaque cellule de classification croisée (i, j) possible et une colonne pour chaque classification par ligne i et chaque classification par colonne j . Donc, la $(i, j)^{\text{e}}$ ligne de $\mu^{\mathbf{xz}}$ contient f_{ij}/N dans les colonnes correspondant à i et à j , et des zéros ailleurs. Donc, θ sera dans l'ensemble des colonnes de $\mu^{\mathbf{xz}}$ si, et uniquement si, pour chaque i et j

Quand la seule unité aberrante est supprimée, laissant 283 724 éléments de la population, les estimateurs à fré-

quences de cellule calées ont d'un peu meilleures propriétés

que l'estimateur sans biais. Pour $n = 200$, l'estimateur à

fréquences de cellules calées linéaire donne une améliora-

tion de 9,3 % de l'efficacité pour l'emploi total et de 16,7 %

pour la rémunération totale. Les ratios comparables pour

l'estimateur à fréquences de cellule calées exponentiel sont

de 10,4 % pour l'emploi total et de 18,0 % pour la rémuné-

ration totale.

Enfin, l'estimateur de variance donné par l'équation (15)

présente un léger biais vers le bas.

thèse d'un modèle de prédiction

du modèle \mathbf{z} , Kott et Chang (2010) émettent plutôt l'hypo-

variables de référence \mathbf{x} est égal au nombre \bar{Q} de variables

non-réponse n'est pas un problème. Si le nombre P de

tiré avec les probabilités de sélection d_k^{-1} de sorte que la

un modèle différent. Dans leur article, S est un échantillon

tiques de $f_{y,z}^{y,z}$ sous une structure probabiliste fondée sur

Kott et Chang (2010) ont établi les priorités asympto-

indépendantes du modèle sous la contrainte

et $I_k = 0$ autrement.

Quand $P > \bar{Q}$, l'équation du modèle (6) doit être

remplacée par

pour une certaine matrice \mathbf{A}_∞ de dimension $\bar{Q} \times P$ [qui est

définie dans un cadre asymptotique approprié, voir Kott et

Chang (2010)].

Donc, quand \mathbf{x} représente les variables indicatrices pour

la classification croisée complète $I \times J$, nous avons que

de la variable de réponse sur la $(i, j)^e$ cellule. Donc, par

fonction de \mathbf{x} , le modèle donné par (6) et (7) est automati-

quement vérifié quand l'échantillonnage (y compris la non-

réponse) est non informatif.

Cependant, dans notre application de calage, $P = IJ >$

$\bar{Q} = I + J - 1$ et il n'existe a priori aucune raison que

l'équation du modèle (8) soit vérifiée.

4. Biais et variance fondés sur un modèle des estimateurs calés

Motivés par Kott et Chang (2010), nous examinons le comportement des estimations calées sous le scénario suivant :

1. Les variables de référence \mathbf{x} sont les variables indi-

catrices pour une partition de la population en classes

C_r . Le modèle (6) est automatiquement vérifié, où la

r^e composante de θ est la moyenne de population dans

de C_r . Soit f_r la proportion de la population dans

C_r et $V_r = \text{Var}(e_k | k \in C_r)$. Nous utiliserons aussi

la notation $\text{Var}(\mathbf{x}_k)$ pour V_r quand $k \in C_r$.

2. L'échantillon est un échantillon aléatoire simple de

taille n sélectionné avec remise.

3. La fonction de lien rétrotransformée $f(\eta)$ dans l'esti-

mateur $f_{y,z}^{y,z}$ de l'équation (2) est $f(\eta) = 1 + \eta$.

Bien que ces hypothèses ne soient pas réalistes en

pratique, l'objectif principal de la présente section est de

justifier de manière heuristique un choix, décrit à la section

suivante, pour la matrice \mathbf{V} . À ce stade, nous n'imposons

aucune contrainte sur \mathbf{z} .

Nous notons que, dans cette situation, $E(e_k | \mathbf{x}_{j^c}, I_{j^c},$

$f \in \mathcal{U}) = 0$. Notons que l'équation (7) sera vérifiée si les

composantes des variables du modèle \mathbf{z} sont des fonctions

linéaires de \mathbf{x} , c'est-à-dire que chaque composante de \mathbf{z} est cons-

istante dans chaque classe. Cependant, si $P > \bar{Q}$, l'équation

(8) ne sera généralement pas vérifiée. De toute façon, à la

présente section, nous n'avons besoin ni de (7) ni de (8).

Nous posons que

$$\mu_{\mathbf{x}} = \frac{1}{N} \sum_{j \in J} \mathbf{x}_{j^c} \quad \mu_{\mathbf{z}} = \frac{1}{N} \sum_{j \in J} \mathbf{z}_{j^c}$$

et la matrice \mathbf{A}_∞ de l'équation (8) devient

$$\mathbf{A}_\infty = \mathbf{V}^{-1} \mu_{\mathbf{z}^c}^{\mathbf{z}^c}$$

Soit $\hat{\mu}_{y,z}^{y,z} = N^{-1} \hat{\mu}_{y,z}^{y,z}$ ou $\hat{\mu}_{y,z}^{y,z}$ est défini comme dans

(2). Nous avons supprimé f dans la notation $\hat{\mu}_{y,z}^{y,z}$ parce

qu'à la présente section, $f(\eta) = 1 + \eta$. En désignant par

$\bar{\mu}_{\mathbf{x}}$ et $\bar{\mu}_{\mathbf{z}}$ les moyennes d'échantillon indiquées et en utili-

sant Kott et Chang (2010), nous avons

$$\hat{\mu}_{y,z}^{y,z} = \bar{\mu}_{\mathbf{x}} + (\bar{\mu}_{\mathbf{z}} - \bar{\mu}_{\mathbf{x}}) \left(\frac{1}{I} \sum_{j \in J} \mathbf{z}_{j^c} \mathbf{x}_{j^c}^T \mathbf{A}_\infty \right) \left(\frac{1}{I} \sum_{j \in J} \mathbf{z}_{j^c} \mathbf{x}_{j^c}^T \right)^{-1} + O_p(n^{-1/2})$$

$$= \bar{\mu}_{\mathbf{x}} + (\bar{\mu}_{\mathbf{z}} - \bar{\mu}_{\mathbf{x}}) \left(\bar{\mathbf{A}}_\infty \mathbf{V} \mathbf{A}_\infty \right)^{-1} \bar{\mu}_{\mathbf{z}}^T \mathbf{A}_\infty + O_p(n^{-1/2})$$

$$= \bar{\mu}_{\mathbf{x}} + (\bar{\mu}_{\mathbf{z}} - \bar{\mu}_{\mathbf{x}}) \mathbf{V}^{-1} \mu_{\mathbf{z}^c}^{\mathbf{z}^c} \theta + O_p(n^{-1/2}). \quad (9)$$

Si $\hat{\mu}_{y,z}^{y,z}$ est bonne, comme cela se produirait si l'on modi-

fie $f(\eta) = 1 + \eta$ quand η est grand afin d'éviter de

grands ajustements des poids de calage, nous avons

Statistique Canada, N° 12-001-X au catalogue

cellule calées exponentiel, l'accroissement de l'efficacité comparativement à l'EQM empirique de l'estimateur non pondéré est de 8,6 % pour l'emploi total et de 13,5 % pour la rémunération totale. La comparaison aux valeurs théoriques pour l'estimateur pondéré penche en faveur des estimateurs à fréquences de cellule calées, mais nous utilisons les résultats empiriques obtenus pour l'estimateur non pondéré, car les mêmes échantillons Monte Carlo ont été employés pour les divers estimateurs. L'estimateur à fréquences de cellule calées exponentiel contient d'être meilleurs que l'estimateur sans biais en ce qui concerne l'EQM pour une taille d'échantillon de $n = 1\,000$.

Tableau 4

Comparaison empirique de quatre estimateurs de l'emploi total : population avec suppression de l'unité aberrante.

Estimateur	Biais	Erreur-type	REQM	R.C. var. est.
Non pondéré (théorique)	0	950 688	975 632	965 448
Non pondéré (empirique)	5 395	975 617	1 019 599	963 314
Marges calées	5 777	1 019 583	1 019 599	963 314
Fréq. cell. calées	-211 568	909 070	933 365	877 343
Ratissage	6 888	1 018 383	1 018 405	956 867
Fréq. cell. calées exponentiel	-217 810	902 756	928 660	868 797
Non pondéré (théorique)	0	424 552	422 116	414 019
Non pondéré (empirique)	-8 393	422 116	422 199	414 019
Marges calées	-9 430	418 153	418 259	408 577
Fréq. cell. calées	-58 808	408 391	412 603	399 961
Ratissage	-61 014	407 780	412 320	399 311
Fréq. cell. calées exponentiel	0	188 517	191 632	188 089
Non pondéré (théorique)	0	188 517	191 632	188 089
Non pondéré (empirique)	702	191 631	191 632	188 089
Poststratifié (théorique)	0	187 691	190 855	187 180
Poststratifié (empirique, 9 cas exclus)	563	190 854	190 855	187 180
Marges calées	820	190 662	190 664	186 664
Fréq. cell. calées	-9 376	189 884	190 115	186 202
Ratissage	2 933	205 924	205 944	186 618
Fréq. cell. calées exponentiel	-9 922	189 813	190 072	186 140

Tableau 5

Comparaison empirique de quatre estimateurs de la rémunération totale : population avec suppression de l'unité aberrante

Estimateur	Biais	Erreur-type	REQM	R.C. var. est.
Non pondéré (théorique)	0	1 330 930	1 341 900	1 334 556
Non pondéré (empirique)	711	1 341 900	1 341 901	1 334 556
Marges calées	1 256	1 387 484	1 387 485	1 318 285
Fréq. cell. calées	-201 575	1 225 852	1 242 314	1 194 071
Ratissage	1 473	1 386 978	1 386 979	1 311 353
Fréq. cell. calées exponentiel	-206 956	1 217 881	1 235 340	1 184 166
Non pondéré (théorique)	0	594 370	587 775	582 524
Non pondéré (empirique)	-8 169	587 775	587 832	582 524
Marges calées	-10 093	583 606	583 693	576 251
Fréq. cell. calées	-56 429	569 158	571 948	563 022
Ratissage	-10 529	584 532	584 626	576 282
Fréq. cell. calées exponentiel	-58 435	568 277	571 273	562 061
Non pondéré (théorique)	0	263 923	266 779	264 110
Non pondéré (empirique)	1 185	266 779	266 782	264 110
Poststratifié (théorique)	0	263 339	265 973	263 210
Poststratifié (empirique, 9 cas exclus)	566	265 973	265 973	263 210
Marges calées	991	265 449	265 451	262 556
Fréq. cell. calées	-8 565	264 126	264 265	261 483
Ratissage	-6 008	271 535	271 602	262 021
Fréq. cell. calées exponentiel	-9 070	264 038	264 194	261 394

Comparaison empirique de quatre estimateurs de la rémunération totale (dizaines de milliers de dollars)

Estimateur	Biais	Erreur-type	REOM	R.C. var. est.
Non pondéré (théorique)	0	1 682 571	1 551 226	1 543 483
Non pondéré (empirique)	-11 119	1 551 186	1 582 425	1 510 413
Marges calées	-11 474	1 582 383	1 467 664	1 413 411
Fréq. cell. calées	-214 323	1 451 931	1 579 882	1 501 170
Ratissage	-221 435	1 438 810	1 455 750	1 393 246
Fréq. cell. calées exponentiel	-221 435	1 438 810	1 455 750	1 393 246
Non pondéré (théorique)	0	751 406	772 501	768 878
Non pondéré (empirique)	-2 911	772 495	776 968	768 869
Marges calées	-4 372	776 955	757 963	751 384
Fréq. cell. calées	-51 649	756 201	778 316	769 428
Ratissage	-4 684	778 302	756 913	749 832
Fréq. cell. calées exponentiel	-54 305	754 963	756 913	749 832
Non pondéré (théorique)	0	333 654	336 068	337 239
Non pondéré (empirique)	2 678	336 057	335 276	336 192
Poststratifié (théorique)	0	333 765	334 920	336 064
Poststratifié (empirique)	1 802	335 271	335 637	335 006
Marges calées	2 510	334 910	339 106	335 230
Fréq. cell. calées	-7 149	333 560	334 493	334 755
Ratissage	-4 679	339 074	334 493	334 755
Fréq. cell. calées exponentiel	-9 251	334 365	334 493	334 755

L'examen des tableaux 2 et 3 révèle que les méthodes pour lesquelles $P > \bar{Q}$, c'est-à-dire celles consistant à caler les fréquences de cellule de la classification croisée en utilisant des poids de calage qui dépendent des classifications marginales, présentent manifestement un biais plus important que les autres. Cependant, relativement aux comparables à ceux des tableaux 2 et 3.

réelle. Les échantillons utilisés pour les tableaux 4 et 5 sont identiques à ceux utilisés pour les tableaux 2 et 3, excepté que, dans les cas où l'unité aberrante était incluse dans l'échantillon, elle a été remplacée par une nouvelle observation provenant de la population. Ce remplacement a été effectué afin de rendre les résultats des tableaux 4 et 5 plus

Les distributions des variables réponse, c'est-à-dire l'emploi total et la rémunération totale, sont fortement asymétriques avec étallement vers la droite. Il existe une entrecroisement (dans l'Etat C et le groupe d'industries 4) dont l'emploi total est égal à plus du double de l'emploi total de l'entreprise suivante par ordre décroissant de taille et à plusieurs centaines de fois l'emploi moyen des entreprises restantes. Nous avons répété l'étude en utilisant une population dont cette entreprise a été supprimée. Les résultats sont présentés aux tableaux 4 et 5. En pratique, pour cette population, l'entreprise en question serait normalement échantillonnée avec certitude (*unité auto-représentative*) et les échantillons seraient constitués en partant des entreprises restantes. Donc, les tableaux 4 et 5 reflètent peut-être davantage les propriétés relatives des estimateurs dans la pratique

Il semble aussi que la fonction de lien rétrotransformée f exponentielle donne de légèrement de meilleurs résultats que le choix linéaire. Du point de vue de l'intensité des calculs, la première est beaucoup plus coûteuse que la seconde. Nous constatons aussi que, quand la taille de l'échantillon augmente, les propriétés des estimateurs semblent converger. Cela n'est pas inattendu, parce qu'en l'absence de non-réponse, quand $n \rightarrow \infty$, $\beta \rightarrow 0$, de sorte que les poids d'ajustement $w = f(\mathbf{z})'\beta \rightarrow 1$.

d'ajustement $w = f(\mathbf{z}^T \beta) \mapsto 1$. Si l'on compare l'estimateur à fréquences de cellule

calées linéaire aux valeurs empirique de l'estimateur non pondéré, quand $n = 200$, le premier est, selon l'EQM, environ 7,3 % plus efficace pour l'emploi total et 11,7 % plus efficace pour la rémunération totale. (Cela signifie, par exemple, que l'EQM empirique de l'estimateur non pondéré est égale à 1,117 fois l'EQM empirique de l'estimateur à fréquences de cellule calées linéaire quand on estime la rémunération totale). Pour l'estimateur à fréquences de

moyenne des variances estimées en utilisant le premier terme de l'équation (15). À titre comparatif, nous présentons les valeurs théoriques et empiriques de l'estimateur non pondéré $N / n \sum_{k \in S} y_k^2$. Ces résultats figurent au tableau 2 pour l'emploi total et au tableau 3 pour la rémunération totale.

Pour l'échantillon de taille $n = 5\,000$, la taille prévue d'échantillon dans la cellule la plus petite (État E et groupe d'industries 5) est de 9,07. Bien que cela soit un peu faible pour la poststratification, la probabilité que cette cellule ait

une taille d'échantillon inférieure à 2, c'est-à-dire la taille minimale nécessaire pour l'estimation de la variance, est de 0,0011. Dans nos simulations, neuf exécutions ont produit une cellule ayant une taille inférieure à 2. Nous présentons l'estimation de variance (7,65) de Särndal, Swensson et Wretman (1992), ainsi que son comportement théorique en utilisant l'approximation de la variance donnée par (7,6) dans Särndal et coll. (1992).

Tableau 1

Entités commerciales selon l'État et le groupe d'industries

Groupe d'industries					
	1	2	3	4	5
A	5 986	5 548	7 712	3 969	1 299
B	(2,11 %)	(1,96 %)	(2,72 %)	(1,40 %)	(0,46 %)
	18 782	31 572	22 012	4 982	4 504
	(6,62 %)	(11,13 %)	(7,76 %)	(1,76 %)	(1,59 %)
C	13 518	13 099	17 837	5 610	3 001
	(4,76 %)	(4,62 %)	(6,29 %)	(1,98 %)	(1,06 %)
D	30 428	36 017	32 541	10 963	5 399
	(10,72 %)	(12,69 %)	(11,47 %)	(3,86 %)	(1,90 %)
E	2 225	2 020	3 110	1 076	515
	(0,78 %)	(0,71 %)	(1,10 %)	(0,38 %)	(0,18 %)
somme	70 939	88 256	83 212	26 600	14 718
	(25,00 %)	(31,11 %)	(29,33 %)	(9,38 %)	(5,19 %)

Tableau 2
Comparaison empirique de quatre estimateurs de l'emploi total

Estimateur	Biais	Erreur-type	REQM	R.C. var. est.
Non pondéré (théorique)	0	1 113 220	1 068 944	1 059 463
Non pondéré (empirique)	-1 280	1 068 944	1 068 945	1 048 873
Marges calées	-1 394	1 105 201	1 105 201	975 140
Fréq. cell. calées	-218 751	1 008 436	1 031 889	1 041 490
Ratissage	-462	1 103 172	1 103 172	962 153
Fréq. cell. calées exponentiel	-227 578	1 000 154	1 025 719	
$n = 200$				
Non pondéré (théorique)	0	497 144	505 970	501 144
Non pondéré (empirique)	-5 435	505 941	506 277	498 946
Marges calées	-6 212	506 239	496 790	488 222
Fréq. cell. calées	-56 118	493 611	507 961	499 237
Ratissage	-4 854	507 938	496 445	487 281
Fréq. cell. calées exponentiel	-58 891	492 939		
$n = 1\,000$				
Non pondéré (théorique)	0	220 751	224 088	222 034
Non pondéré (empirique)	1 516	224 088	224 093	
Poststratifié (théorique)	0	220 315		
Poststratifié (empirique, 9 cas exclus)	1 234	223 225	223 228	221 094
Marges calées	1 649	223 091	223 098	220 833
Fréq. cell. calées	-8 606	222 170	222 337	220 347
Ratissage	3 632	236 355	236 383	220 606
Fréq. cell. calées exponentiel	-10 643	223 472	223 725	220 207
$n = 5\,000$				

Les mêmes formules marchent pour la non-couverture, auquel cas $f(\mathbf{z})^T \mathbf{b}_0^{-1}$ est la probabilité combinée de réponse et de couverture.

Nous désignons par N_{ij}^y , S_{ij}^y , et r_{ij}^y la taille de la population, l'échantillon de répondants et la taille de l'échantillon de répondants dans la cellule (i, j) . Bien que nous supposions que N_{ij}^y est connu, notre méthodologie n'exige pas que les classifications par ligne et par colonne des non-répondants soient connues.

Nous définissons $N_i^y = \sum_j N_{ij}^y$ et, de manière analogue, nous définissons N_j^y .

Nous utiliserons les estimateurs d'un total T_y de la forme

$$t_y = \frac{n}{N} \sum_{j \in S_y} \sum_{i \in S_j} \frac{1}{w_{ij}^y} y_k \quad (4)$$

où les poids d'ajustement w_{ij}^y sont définis comme il est indiqué plus bas. Ces estimateurs sont tous des cas particuliers des équations (2) et (3) quand nous utilisons $\mathbf{V} = \mathbf{I}$.

Dans l'estimateur à marges calées, nous utilisons $f(\mathbf{n}) = \mathbf{I} + \boldsymbol{\eta}$ et définissons $\mathbf{x} = \mathbf{z}$ comme étant $I + J - 1$ variables indicatrices indépendantes pour les catégories de marge. Dans ce cas, T_x est un vecteur de N_i^y et N_j^y . Les poids d'ajustement $f(\mathbf{z})^T \mathbf{b}$ sont de la forme $w_{ij}^y = 1 + \beta_i + \beta_j$ quand \mathbf{z} est le vecteur des variables indicatrices de l'appartenance à la (i, j) classification par ligne et colonne, respectivement. Puisque le nombre d'équations (la dimension de \mathbf{x}) est égal au nombre d'inconnues (la dimension de β), nous devrions être capables de résoudre exactement les équations

$$T_x = \sum_{k \in S} d_k^x f(\mathbf{z})^T \mathbf{b} \mathbf{x}_k. \quad (5)$$

Donc les β_i, β_j résolvent les équations linéaires de rang $I + J - 1$

$$N_i^y = \frac{n}{N} \sum_j (1 + \beta_i + \beta_j) r_{ij}^y$$

$$N_j^y = \frac{n}{N} \sum_i (1 + \beta_i + \beta_j) r_{ij}^y$$

ce qui découle facilement de (5).

Dans l'estimateur à fréquences de cellule calées, nous utilisons $f(\mathbf{n}) = \mathbf{I} + \boldsymbol{\eta}$ et définissons \mathbf{x} comme étant les IJ variables indicatrices pour la classification croisée complète et \mathbf{z} comme étant les $I + J - 1$ variables indicatrices indépendantes pour les catégories de marge. Dans ce cas, T_x est un vecteur de N_{ij}^y et, puisque $\mathbf{V} = \mathbf{I}$, les poids d'ajustement $w_{ij}^y = 1 + \beta_i + \beta_j$ minimisent la fonction d'objectif

$$\left[\sum_{i,j} \sum_{k \in S_{ij}} N_{ij}^y - \frac{n}{N} \sum_j \sum_i (1 + \beta_i + \beta_j) r_{ij}^y \right]^2.$$

3. Étude empirique

Nous utilisons comme population les données provenant du Quarterly Census of Employment et Wages (QCEW) de recueilles pour le premier trimestre de 2005, en nous limitant à cinq États et cinq groupes d'industries. Le QCEW est compilé d'après des rapports qui doivent obligatoirement être soumis aux bureaux de l'emploi des États, de sorte qu'il s'agit pratiquement d'un recensement, et nous avons utilisé les données du QCEW complet pour les cinq États et les cinq groupes d'industries choisis. Cette population compte $N = 283\,725$ entreprises réparties comme il est indiqué au tableau 1.

Les variables réponse y sont l'emploi total et la rémunération (trimestrielle) totale. Pour ces variables $T_y = 2\,981\,364$ pour l'emploi total et $T_y = 2\,334\,400$ (en dollars de milliers de dollars) pour la rémunération totale. Dans la présente étude, nous avons tiré 10 000 échantillons de taille $n = 200,1\,000$ et 5 000. Pour chacun des quatre estimateurs, nous présentons des estimations du biais, de l'erreur-type et de la racine carrée de l'erreur quadratique moyenne. Nous donnons aussi la racine carrée de la

$$\left[\sum_{i,j} \sum_{k \in S_{ij}} N_{ij}^y - \frac{n}{N} \sum_j \sum_i \exp(\beta_i + \beta_j) r_{ij}^y \right]^2.$$

Chang et Kott (2008) donnent les formules de l'estimation de la variance de t_y , fondée sur l'échantillon. À l'anneau, nous appliquons ces formules aux quatre estimateurs susmentionnés.

longueur P du vecteur de variables de référence \mathbf{x} , f est une fonction à valeur réelle positive que Chang et Kott (2008) appellent *back link function* (fonction de *lien rétro-transformée* ou fonction de *lien inverse*), et V est une matrice définie positive symétrique de dimensions $P \times P$. V peut dépendre de β_j , comme cela se produirait si $(V(\beta))$ était une mesure de la variabilité de $\sum_{k \in S} d_k f(\mathbf{z}_k^T \beta) \mathbf{x}_k^T$.

Dans Chang et Kott (2008), l'échantillon réalisé S est constitué des répondants d'un échantillon original auquel sont associés les poids d'échantillonnage d_k . L'échantillon de répondants S est supposé être un sous-échantillon de Poisson de l'échantillon original dont les probabilités de sélection poissonniennes sont $f(\mathbf{z}_k^T \beta_0)^{-1}$, pour un certain β_0 . Les formules asymptotiques dérivées par Chang et Kott l'ont été sous un cadre asymptotique pour ce modèle de quasi-randomisation (fondé sur le plan). Nous utilisons le terme *quasi-randomisation* afin de nous rappeler que le mécanisme de réponse de Poisson supposé est, en fait, fondé sur un modèle.

Il convient de souligner que l'utilisation du calage pour corriger la non-réponse remonte à Fuller, Loughin et Baker (1994), du moins quand $\mathbf{z} = \mathbf{x}$ et $f(\eta) = 1 + \eta$.

Nous proposons d'utiliser la méthodologie de Chang et Kott (2008) en gardant \mathbf{x} comme variables indicatrices pour la classification croisée $I \times J$ complète, mais en supposant que \mathbf{z} est un vecteur de $I + J - 1$ variables indicatrices pour les classifications marginales. Autrement dit, nous proposons de rééquilibrer l'échantillon afin de nous rapprocher autant que possible, au sens de la minimisation de (3), des proportions de cellule correctes dans la classification complète, mais en exigeant que les poids de rééquilibrage ne dépendent que des classifications marginales.

Le cadre de Chang et Kott (2008) est applicable en présence de non-réponse (et/ou de non-couverture) si $f(\mathbf{z}_k^T \beta_0)^{-1}$ est la probabilité de réponse (ou de réponse et de couverture) combinée. Nous notons que la poststratification, qui est un cas particulier de calage, est souvent utilisée en vue de corriger la non-réponse ou la non-couverture. Dans l'exemple test que nous présentons plus bas, aucune non-réponse ni non-couverture ne doit être corrigée, et le cadre de Chang et Kott (2008) s'applique donc avec $\beta_0 = \mathbf{0}$ pour toute fonction f avec $f(0) = 1$. Autrement dit, si le calage est utilisé uniquement pour rééquilibrer l'échantillon, nous pouvons utiliser le cadre de Chang et Kott (2008) avec pratiquement n'importe quelle fonction f . Mais si nous essayons de corriger la non-réponse et/ou la non-couverture, des hypothèses plus fortes sont nécessaires.

Il convient de souligner que le ratissage (*raking*) est simplement l'estimation calée en utilisant les $I + J - 1$ variables indicatrices des classifications marginales comme variables de référence ainsi que variables du modèle, et en

2. Formules mathématiques

À la présente section, nous donnons les formules utilisées dans l'étude. Toutes sont des cas particuliers des formules présentées dans Chang et Kott (2008). Nous supposons qu'un échantillon aléatoire simple de taille n est tiré d'une population de taille N et nous utilisons S et r pour désigner l'échantillon des répondants et sa taille. Nous supposons que la fonction des poids de calage possède un β_0 tel que $f(\mathbf{z}^T \beta_0)^{-1}$ est la probabilité de réponse pour un élément ayant les variables du modèle \mathbf{z} . En particulier, et sans perte de généralité, s'il n'existe aucun problème de non-réponse, nous supposons que $f(0) = 1$.

À la présente section, nous examinons également l'utilisation de cette fonction de lien rétrotransformée. À la section 2, nous présentons les formules exactes pour les estimateurs que nous utiliserons dans la présente étude. Le cadre de Chang et Kott (2008) peut être appliqué pour établir les estimateurs de variance fondés sur l'échantillon et ces calculs sont exposés en annexe.

À la section 3, nous donnons les résultats d'une étude empirique réalisée en utilisant les données du Quarterly Census of Employment and Wages pour le premier trimestre de 2005 recueillies par le Bureau of Labor Statistics. Nous nous limiterons à cinq États que nous désignerons par A, B, C, D, E et à cinq groupes d'industries que nous désignerons par 1, 2, 3, 4, 5. Nous n'identifierons pas de manière plus précise les États ni les groupes d'industries afin d'empêcher l'identification de l'unité aberrante dans la discussion qui suit. Cette population comprend 283 725 entreprises. De cette population, nous tirons par une méthode de Monte Carlo des échantillons aléatoires simples de taille $n = 200$, 1 000 et 5 000, et procéderons à la classification croisée de l'État par le groupe d'industries.

Il convient de souligner que 0,18 % de la population correspond à la classification croisée de l'État E et du groupe d'industries 5. Donc, quand $n = 200$, la taille prévue d'échantillon dans cette cellule est de 0,36, de sorte qu'il est hors de question de procéder à la poststratification en se basant sur la classification croisée.

Kott et Chang (2010) dérivent les propriétés de $f_{y,z^{IV}}$ en utilisant un cadre fondé sur un modèle. Les modèles qu'ils prennent en considération ne s'appliquent pas à notre choix des variables \mathbf{x} et \mathbf{z} . Cependant, motivés par leur approche, nous examinons à la section 4 le comportement de l'estimateur $f_{y,z^{IV}}$ défini par l'équation (2), sous des hypothèses très simplifiées, y compris $f(\eta) = 1 + \eta$. Cela nous mène, à la section 5, au choix d'une nouvelle matrice de pondération V^{-1} à utiliser dans (3). Nous poursuivons ensuite notre exploration empirique en utilisant ce nouvel estimateur.

Options de calage pour remplacer la poststratification de données croisées à deux dimensions

Ted Chang¹

Résumé

Nous considérons diverses options de remplacement de la poststratification dans le cas de données croisées à deux dimensions pour lesquelles la fréquence d'au moins l'une des cellules du tableau à double entrée est trop faible pour permettre la poststratification en se basant sur cette classification double. Dans le jeu de données utilisé pour l'étude, la fréquence prévue dans la cellule la plus petite est de 0,36. Une approche consiste simplement à regrouper certaines cellules. Néanmoins, il est probable que cela détruise la structure de la classification double. Les approches de remplacement que nous proposons permettent de maintenir la classification à double entrée originale des données. Elles sont fondées sur l'étude du calage de Chang et Kott (2008). Nous choisissons des ajustements des pondérations qui dépendent des classifications marginales (mais non de la classification croisée complète) pour minimiser une fonction d'objectif correspondant aux différences entre les fréquences dans la population et dans les cellules du tableau à double entrée et leurs estimations sur échantillon. Selon la terminologie de Chang et Kott (2008), si les classifications par ligne et par colonne comprennent I et J cellules, respectivement, on obtient IJ variables de *référence* et $I + J - 1$ variables du *modèle*. Nous étudions les propriétés de ces estimateurs en créant des échantillons aléatoires simples pour la simulation d'après les données du Quarterly Census of Employment and Wages de 2005 tenues à jour par le Bureau of Labor Statistics. Nous procédons à la classification croisée de l'État et du groupe d'industries. Dans notre étude, comparativement à l'estimateur FEM pouvant aller jusqu'à 20 % pour un petit échantillon.

Mots clés : Calage ; poststratification ; modèle de prédiction.

1. Introduction

Supposons que nous ayons une population \mathcal{U} doublement stratifiée par deux variables catégoriques désignées par les indices (i, j) , $i = 1, \dots, I$, $j = 1, \dots, J$ et écrivons \mathcal{U}_{ij} pour la strate (i, j) . Si nous tirons un échantillon aléatoire simple S de taille n et que nous désignons par y la variable d'intérêt, un estimateur naturel du total $T_y = \sum_{k \in \mathcal{U}} y_k$ est l'estimateur poststratifié

$$\hat{t}_{y,ps} = \sum_{j=1}^J N_j \bar{y}_j \quad (1)$$

où N_j est la taille de la population \mathcal{U}_j et \bar{y}_j est la moyenne d'échantillon de y sur $S \cap \mathcal{U}_j$. Cet estimateur est généré par un échantillon n_j de $S \cap \mathcal{U}_j$ pour chaque j . Que faire si certaines tailles n_j sont petites, ou même nulles ?

L'approche classique consiste à regrouper certaines cellules jusqu'à ce que toutes les tailles n_j soient suffisamment grandes. Cependant, il pourrait être impossible d'effectuer ce groupement d'une façon qui maintienne le schéma de classification croisée ; autrement dit, les indices j pourraient dépendre de i . L'estimateur poststratifié $\hat{t}_{y,ps}$ est un cas particulier d'un estimateur par calage. Définissons pour chaque $k \in \mathcal{U}$ le

Ces équations établissent que si l'on utilise les variables \mathbf{x} de *référence*, alors $\hat{t}_{y,ps}$ est l'estimateur par calage résultant de T_y . Chang et Kott (2008) ont déterminé les propriétés asymptotiques de l'estimation calée de la forme

$$(2) \quad \hat{t}_{y,z,v} = \sum_{k \in S} d_k f(\mathbf{z}_k^T \beta) y_k$$

où β minimise une fonction d'objectif de la forme

$$(3) \quad \hat{Q}(\beta) = \left(T^x - \sum_{k \in S} d_k f(\mathbf{z}_k^T \beta) \mathbf{x}_k \right)^T V^{-1} \left(T^x - \sum_{k \in S} d_k f(\mathbf{z}_k^T \beta) \mathbf{x}_k \right).$$

Dans les équations (2) et (3), \mathbf{z} est un vecteur de variables du *modèle* dont la longueur Q est au plus égale à la

Longford, N.T. (2006). Calcul de la taille de l'échantillon pour l'estimation pour petits domaines. *Techniques d'enquête*, 32, 97-106.

North American Industry Classification System, Version 1.4 (2008). Catalogue I2F0074XCB, Statistics Canada.

Tableau 5
CV (%) pour les estimateurs composites en utilisant la répartition de Longford : $G = 0$ et $q = 0, 0,5, 1,0$ et $1,5$

Province	$q = 0$	$q = 0,5$	$q = 1,0$	$q = 1,5$
T.-N.	12,7	17,0	24,2	37,3
I.-P.-É.	12,4	23,8	46,0	112,2
N.-B.	10,4	12,8	16,1	20,4
N.-É.	9,4	11,9	14,5	11,7
Qc	10,3	9,0	8,3	8,0
Ont.	13,9	11,1	9,3	8,2
Man.	11,2	13,1	16,0	20,3
Sask.	12,4	14,6	17,9	23,2
Alb.	11,4	11,2	11,5	12,2
C.-B.	14,4	13,3	12,9	13,1
CA	8,0	6,3	5,4	5,6

5.3 Moyennes de domaine

Dans le Registre des entreprises du Canada, les établissements sont classés selon l'industrie en utilisant le Système de classification des industries de l'Amérique du Nord (SCIAN). Ce dernier est avant tout un système de classification des établissements destiné à la compilation de statistiques sur la production. L'industrie associée à chaque établissement inscrit dans le Registre des entreprises du Canada est désignée par un code à six chiffres conforme au SCIAN. En tout, 67 codes à six chiffres sont associés au secteur du commerce de détail. Ces codes sont regroupés en 19 groupes de commerce (GC) aux fins de la publication des données.

Nous avons considéré les groupes de commerce comme des domaines recoupant les provinces (states). Le groupe de commerce comprenant le plus petit nombre d'établissements est le CG 110 (magasins de bière, de vin et de spiritueux) qui compte 307 établissements et le groupe de commerce comprenant le plus grand nombre d'établissements est le CG 100 (dépanneurs et magasins d'alimentation spécialisés) qui regroupe 7 752 établissements. Les établissements ont été codés en fonction des 19 groupes de commerce dans toutes les provinces sauf une, ceux de l'Île-du-Prince-Édouard ayant été codés en ne considérant que 16 groupes de commerce seulement.

Nous avons appliqué la procédure NLP fondée sur (4.5), (4.6) et (4.7), et obtenu l'augmentation de la taille totale optimale d'échantillon requise pour remplir les exigences de fiabilité spécifiée pour les estimateurs de domaine \hat{y}_h^* . Nous avons constaté qu'aucune augmentation de la taille totale d'échantillon n'est nécessaire si la tolérance appliquée à $CV(\hat{y}_h^*)$ est inférieure ou égale à 30 % pour chaque domaine. Si la tolérance est réduite à 25 %, l'augmentation de la taille totale optimale de l'échantillon est de 622, ce qui donne une taille totale d'échantillon après l'augmentation de 4 068. Si la tolérance est abaissée à 20 %, l'augmentation de

6. Résumé et conclusion

la taille totale optimale d'échantillon est de 2 100 et la taille totale de l'échantillon après l'augmentation est de 5 546, c'est-à-dire une valeur considérablement plus grande que la taille originale de 3 446. Il convient de souligner les CV des moyennes de strate \bar{y}_h et de la moyenne d'échantillon pondérée \bar{y}_{st}^* diminuent à mesure que la taille totale d'échantillon augmente.

Nous avons proposé une méthode par programmation non linéaire (NLP) de répartition de l'échantillon entre les strates sous échantillonnage aléatoire stratifié. Cette méthode minimise la taille totale de l'échantillon sous la contrainte des tolérances spécifiées pour les coefficients de variation des estimateurs des moyennes de strate et de la moyenne de population. Nous avons considéré des estimateurs directs ainsi que des estimateurs composites des moyennes de strate. Le cas de domaines recoupant les strates est également étudié. Les difficultés que posent d'autres méthodes en vue de satisfaire aux contraintes de fiabilité spécifiées sont illustrées au moyen de données provenant de l'Enquête mensuelle sur le commerce de détail auprès d'établissements uniques menée par Statistique Canada. Nous constatons aussi que la méthode NLP peut être étendue facilement en vue de traiter des contraintes de fiabilité pour des variables multiples. Les répartitions intermédiaires qui donnent d'assez bons résultats en ce qui concerne les contraintes de fiabilité sont également mentionnées.

Remerciements

Les auteurs remercient deux examinateurs et un rédacteur associé pour leurs commentaires constructifs et suggestions.

Bibliographie

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête*, 15, 49-60.
Bankier, M. (1988). Power allocation: Determining sample sizes for sub-national areas. *The American Statistician*, 42, 174-177.
Cochran, W.G. (1977). *Sampling Techniques*, 3^e Edition. New York : John Wiley & Sons, Inc.
Costa, A., Satorra, A. et Ventura, E. (2004). Using composite estimators to improve both domain and total area estimation. *SORT*, 28, 69-86.
Huddleston, H.F., Claypool, P.L. et Hocking, R.R. (1970). Optimum allocation to strata using convex programming. *Applied Statistics*, 19, 273-278.
Statistique Canada, N° 12-001-X au catalogue

Province		$q = 0$		$q = 0,5$		$q = 1,0$		$q = 1,5$	
T.-N.	13,5	G = 0	19,3	G = 0	23,0	G = 0	22,7	G = 0	30,4
L.-P.-É.	12,0	G = 10	20,4	G = 100	21,4	G = 10	34,0	G = 100	38,3
N.-B.	12,7		25,0		19,4		18,3		29,0
Qc	11,0		9,8		9,4		9,2		8,9
N.-É.	11,1		16,7		14,2		18,7		30,9
Man.	12,7		17,6		14,7		17,7		26,5
Sask.	13,6		18,9		15,7		19,0		28,3
Alb.	13,5		15,7		15,2		13,6		15,9
C.-B.	15,7		16,1		14,7		14,3		15,1
CA	7,3		16,1		6,2		5,5		5,1

Tableau 4 CV (%) pour la répartition de Longford pour $q = 0, 0,5, 1,0$ et $1,5$

5.2 Moyennes de strate : estimateurs composites

Nous présentons maintenant certains résultats pour les estimateurs composites, $\hat{\theta}_h$, des moyennes de strate. Nous avons obtenu la taille totale optimale d'échantillon égale à $n = 3\ 368$ en utilisant la procédure NLP et les exigences de fiabilité (3,6). Cette valeur est un peu plus faible que la valeur optimale $n^0 = 3\ 446$ obtenue pour les estimateurs

Nous examinons maintenant la répartition de Longford (2,4) qui dépend de q et de G . Le tableau 4 donne les résultats pour $q = 0, 0,5, 1,0, 1,5$ et $G = 0, 10, 100$, en utilisant $n = 3\ 446$ obtenu par la méthode NLP. Pour $q = 2,0$, la répartition de Longford ne dépend pas de G et, en fait, se réduit à la répartition de Neyman (1,2) qui minimise $CV(\bar{Y}^{st})$ pour une valeur n fixe, mais donne des valeurs très élevées de $CV(\bar{Y}_h)$, variant de 16 % à 85 % pour sept provinces. L'examen du tableau 4 montre que $CV(\bar{Y}_h)$, pour une valeur donnée de q , augmente rapidement avec G , tandis que $CV(\bar{Y}^{st})$ diminue lentement à mesure que G augmente et, en fait, est presque constant ($\approx 5,1\%$) pour $G > 100$ (valeurs non présentées ici). En outre, $CV(\bar{Y}_h)$, pour une valeur donnée de G , augmente rapidement à mesure que q augmente, tandis que $CV(\bar{Y}^{st})$ diminue. La répartition de Longford pour $q \geq 0,5$ et/ou $G \geq 10$, produit des valeurs de $CV(\bar{Y}_h)$ significativement plus élevées que la tolérance spécifiée $CV_{0h} = 15\%$ pour plusieurs provinces, même si $CV(\bar{Y}^{st})$ respecte la tolérance spécifiée de 6 %. Par ailleurs, pour $q = 0$ et $G = 0$, $CV(\bar{Y}_h)$ est inférieure à la tolérance spécifiée, sauf pour la Colombie-Britannique pour laquelle il est de 15,7 %, mais $CV(\bar{Y}^{st}) = 7,3\%$ et est significativement supérieur à la tolérance spécifiée. Pour $q = 1,0$ et $q = 1,5$, $CV(\bar{Y}_h)$ dépasse 15 % inférieurement à 6 % quand $G = 0$, mais $CV(\bar{Y}_h)$ dépasse 15 % pour six provinces, variant de 17,7 % à 34,0 % pour $q = 1,0$ et de 22,0 % à 54,6 % pour $q = 1,5$. Dans l'ensemble, le tableau 4 donne à penser qu'il n'est possible de trouver aucune combinaison appropriée de q et G telle que toutes les exigences de fiabilité spécifiées soient remplies, même approximativement.

5.2 Moyennes de strate : estimateurs composites

Nous présentons maintenant certains résultats pour les estimateurs composites, $\hat{\theta}_h$, des moyennes de strate. Nous avons obtenu la taille totale optimale d'échantillon égale à $n = 3\ 368$ en utilisant la procédure NLP et les exigences de fiabilité (3,6). Cette valeur est un peu plus faible que la valeur optimale $n^0 = 3\ 446$ obtenue pour les estimateurs

Dans le cas de la répartition de Longford, nous avons utilisé $n = 3\ 368$ et calculé la répartition de l'échantillon et les CV associés des estimateurs composites $\hat{\theta}_h$ et de la moyenne pondérée \bar{Y}^{st} pour des valeurs spécifiées de q et G , en contraignant n_h d'être au moins égal à deux. Pour les données de l'EMCD que nous avons utilisées, le premier terme de (3,5) est petit comparativement au deuxième. Par conséquent, la répartition de l'échantillon est uniforme pour les diverses valeurs de G pour une valeur donnée de q , ce qui signifie que, pour la répartition de Longford, les CV ne varient pas de manière significative en fonction de G .

Par conséquent, nous ne présentons au tableau 5 que les résultats pour $G = 0$ et $q = 0, 0,5, 1,0$ et $1,5$. Nous constatons, en examinant ce tableau, que $CV(\hat{\theta}_h)$ diminue en même temps que q pour les deux plus grandes provinces (Québec et Ontario) parce que l'échantillon se déplace des provinces plus petites vers ces deux provinces à mesure que q augmente. En outre, $CV(\hat{\theta}_h)$ diminue au départ pour l'Alberta et la Colombie-Britannique, mais commence à augmenter quand q est grand, parce que l'échantillon commence à se déplacer de ces provinces vers le Québec et l'Ontario également. En outre, $CV(\hat{\theta}_h)$ augmente avec q pour toutes les autres provinces, sauf la Nouvelle-Écosse pour laquelle il commence à diminuer quand la valeur de q est grande, en raison d'une composante synthétique plus importante et d'un biais très négligeable. En particulier, $CV(\hat{\theta}_h)$ augmente rapidement pour Terre-Neuve et l'Île-du-Prince-Édouard à cause du biais très important.

Par ailleurs, $CV(\bar{Y}^{st})$ diminue au départ quand q augmente, mais commence à augmenter quand q est grand parce que la majeure partie de l'échantillon est affectée au Québec et à l'Ontario et qu'une très petite part est affectée aux provinces plus petites. Le tableau 5 suggère que la répartition de Longford ne donne d'assez bons résultats que pour $q = 0$ et $G = 0$, les valeurs de $CV(\hat{\theta}_h)$ étant inférieures à 15 % pour toutes les provinces, au prix d'une valeur de $CV(\bar{Y}^{st})$ égale à 8,0 %.

spécifiée $CV_0 = 15\%$ pour deux des grandes provinces (QC : 11,4 % et Ont. : 11,0 %) et atteint un CV de 15 % pour les autres provinces.

En utilisant la taille globale optimale d'échantillon égale à 3 446, nous avons calculé les répartitions de l'échantillon n_h et les coefficients de variation $CV(\bar{y}_h)$ et $CV(\bar{y}_{st})$ associés pour l'équitépartition modifiée, la répartition proportionnelle et la répartition fondée sur la racine carrée, qui sont présentées au tableau 2. Ce dernier montre clairement que l'équitépartition modifiée ne convient pas pour ce qui est de satisfaire aux tolérances spécifiées pour les CV, car elle donne $CV(\bar{y}_{st}) = 9,1\%$, valeur significativement plus grande que la valeur spécifiée $CV_0 = 6\%$. En outre, sous l'équitépartition modifiée, $CV(\bar{y}_h)$ égale 19,3 %, 16,3 % et 16,2 % pour les grandes provinces que sont l'Ontario, l'Alberta et la Colombie-Britannique, respectivement. Notons que pour l'Île-du-Prince-Édouard, qui est la plus petite province, le tableau 2 donne $CV(\bar{y}_h) = 0\%$ pour l'équitépartition modifiée, car pour cette province, la méthode donne $n_h = N_h$.

Tableau 1
Valeurs de population pour l'EMCD

Provinces	N_h	\bar{Y}_h	S_h	C_h
Terre-Neuve (T.-N.)	909	963	1 943	2,02
Île-du-Prince-Édouard (Î.-P.-É.)	280	712	1 375	1,93
Nouveau-Brunswick (N.-B.)	1 333	1 368	3 200	2,34
Nouvelle-Écosse (N.-É.)	1 153	1 568	4 302	2,74
Québec (Qc)	11 135	2 006	4 729	2,36
Ontario (Ont.)	21 531	1 722	6 297	3,66
Manitoba (Man.)	1 700	1 295	2 973	2,30
Saskatchewan (Sask.)	1 743	1 212	3 019	2,49
Alberta (Alb.)	5 292	1 698	5 358	3,16
Colombie-Britannique (C.-B.)	7 803	1 291	4 013	3,11
Canada (CA)	52 879	1 654	-	-

Tableau 2
Équitépartition et répartitions proportionnelle, fondée sur la racine carrée et NLP, et CV (%) associés

Province	Équitépartition	Proportionnelle	Racine carrée	NLP
n_h	CV_h	n_h	CV_h	n_h
T.-N.	352	8,4	59	25,4
Î.-P.-É.	280	0,0	18	44,1
N.-B.	352	10,7	87	24,2
N.-É.	352	12,2	75	30,6
Qc	352	12,4	726	8,5
Ont.	352	19,3	1 403	9,4
Man.	352	10,9	111	23,2
Sask.	352	11,9	114	22,6
Alb.	352	16,3	345	16,4
C.-B.	352	16,2	508	13,3
CA	3 446	9,1	3 446	5,2

Dans le cas de la répartition proportionnelle, le tableau 2 donne $CV(\bar{y}_{st}) = 5,2\%$, mais cette méthode donne des CV spécifiés de 15 % pour sept provinces, les chiffres variant de 16,4 % à 44,1 %. Par ailleurs, le tableau 2 montre que la répartition fondée sur la racine carrée offre un compromis raisonnable pour ce qui est du respect des tolérances spécifiées pour les CV. Nous avons $CV(\bar{y}_{st}) = 6,3\%$ et $CV(\bar{y}_h) \leq 15\%$ pour sept provinces, les trois provinces pour lesquelles le CV est supérieur à 15 % étant la Saskatchewan avec 15,2 %, l'Île-du-Prince-Édouard avec 16,2 % et la Nouvelle-Écosse avec 18,1 %.

Le tableau 3 donne les résultats pour la répartition de Costa et coll. (2.1) pour $k = 0,25$, 0,50 et 0,75, en utilisant la valeur $n = 3 446$ obtenue par programmation non linéaire (NLP). Nous constatons d'après le tableau 2 que le choix $k = 0,25$, qui donne plus de poids à l'équitépartition, n'est pas satisfaisant pour l'estimation de la moyenne de la population (Canada), le coefficient de variation étant $CV(\bar{y}_{st}) = 7,2\%$, mais donne de bons résultats pour les moyennes de strate, sauf dans le cas de l'Alberta pour laquelle $CV(\bar{y}_h) = 16,3\%$. Par ailleurs, le choix $k = 0,75$, qui accorde plus de poids à la répartition proportionnelle, donne des résultats médiocres pour l'estimation des moyennes provinciales, $CV(\bar{y}_h)$ variant de 16,2 % à 21,4 % pour sept provinces, quoiqu'il que $CV(\bar{y}_{st})$ soit inférieur à la tolérance souhaitée de 6 %. Le choix intermédiaire $k = 0,50$ donne d'assez bons résultats, les coefficients de variation étant $CV(\bar{y}_{st}) = 6,4\%$ et $CV(\bar{y}_h)$ environ égal à 15 % ou moins, sauf pour deux provinces (Nouvelle-Écosse et Alberta) pour lesquelles les CV sont de 17,0 % et 16,5 %, respectivement. Les propriétés de la méthode de Costa et coll. pour $k = 0,50$ et de la répartition fondée sur la racine carrée sont plus ou moins semblables, et ni l'une ni l'autre ne dépend de la variable d'intérêt, y, contrairement aux répartitions de Longford et NLP.

Tableau 3
Répartition de Costa et coll. et CV (%) associés pour $k = 0,25$, 0,50 et 0,75

Province	$k = 0,25$	$k = 0,50$	$k = 0,75$
n_d	CV_d	n_d	CV_d
T.-N.	278	10,1	205
Î.-P.-É.	214	6,4	149
N.-B.	286	12,3	219
N.-É.	282	14,2	213
Qc	446	10,9	539
Ont.	615	14,5	878
Man.	292	12,2	231
Sask.	292	13,3	733
Alb.	350	16,3	349
C.-B.	391	15,3	430
CA	3 446	7,2	3 446

et (2.8), où $REQM(\theta_h) = EQM(\theta_h)/\sqrt{2}$ et $REQM_{0h}$ est une tolérance spécifiée. L'approximation (3) de $EQM(\theta_h)$ est utilisée dans (3.6).

4. Répartition pour l'estimation sur domaine

Supposons que la population U est partitionnée en domaines dU ($d = 1, \dots, D$) qui recoupent les strates. En

outre, supposons que les estimateurs des moyennes de domaine doivent satisfaire à des tolérances de variance relative spécifiées, dVR_0 , $d = 1, \dots, D$. Nous utilisons la méthode NLP pour trouver les tailles additionnelles optimales d'échantillon de strate nécessaires pour satisfaire aux tolérances de domaine.

Un estimateur de la moyenne de domaine ${}^d\bar{Y} = {}^dN^{-1} \sum_{k \in U_d} y_k$ est l'estimateur par le ratio

$${}^d\hat{\bar{Y}} = \frac{\sum_{h=1}^L N_h n_h^{-1} \sum_{k \in s_h} {}^d\delta_k y_k}{\sum_{h=1}^L N_h n_h^{-1} \sum_{k \in s_h} {}^d\delta_k} \quad (4.1)$$

où ${}^d\delta_k = 1$ si $k \in {}^dU$ et ${}^d\delta_k = 0$ autrement, s_h est l'échantillon tiré de la strate h et dN est la taille du domaine d . La variance relative de l'estimateur par le ratio (4.1) est $VR({}^d\hat{\bar{Y}}) / ({}^d\bar{Y})^2$, où la variance $V({}^d\hat{\bar{Y}})$ s'obtient au moyen de la formule de linéarisation habituelle pour un estimateur par le ratio.

Soit n_h la taille totale révisée de l'échantillon tiré de la strate h , telle que l'accroissement de l'échantillon provenant de la strate h est $n_h - n_0$. Soit $\tilde{f}_h = n_h / N_h$ la fraction d'échantillonnage correspondante. Nous obtenons la taille optimale $\tilde{n} = (n_1, \dots, n_L)^T$ en minimisant l'accroissement de l'échantillon

$$g(\tilde{\mathbf{f}}) = -1 + \sum_{d=1}^D \sum_{h=1}^L ({}^d\tilde{f}_h - {}^d\tilde{f}_0) N_h \quad (4.2)$$

par rapport à $\tilde{\mathbf{f}} = ({}^d\tilde{f}_1, \dots, {}^d\tilde{f}_L)^T$ sous les contraintes

$${}^d\tilde{f}_0 \leq {}^d\tilde{f}_h \leq 1, \quad h = 1, \dots, L \quad (4.3)$$

$$VR({}^d\hat{\bar{Y}}) \leq {}^dVR_0, \quad d = 1, \dots, D. \quad (4.4)$$

Comme précédemment, nous reformulons le problème en exprimant (4.2), (4.3) et (4.4) en fonction de $\tilde{\mathbf{k}} = (k_1, \dots, k_L)^T$, où $k_h = \tilde{f}_h^{-1}$. Cela nous mène à la minimisation de la fonction convexe séparable

$$g^*(\tilde{\mathbf{k}}) = \sum_{h=1}^L N_h \tilde{k}_h^{-1} \quad (4.5)$$

$$VR({}^d\hat{\bar{Y}}) = \left(\frac{{}^d\bar{Y}^{-2} \sum_{h=1}^L \left(\frac{{}^dN}{N_h} \right) \left(\frac{{}^dN}{2\tilde{k}_h} - 1 \right) S_{ch}^2 \leq {}^dVR_0, \quad d = 1, \dots, D \right) \quad (4.7)$$

$$1 \leq \tilde{k}_h \leq k_0, \quad h = 1, \dots, L \quad (4.6)$$

par rapport à $\tilde{\mathbf{k}}$ sous les contraintes linéaires

où dVR_0 est la tolérance spécifiée, ${}^dS_{ch}^2$ désigne la variance de strate des résidus ${}^d\delta_k(y_k - {}^d\bar{Y})$ pour $k \in U_h$ et U_h désigne la population de la strate. Désignons les valeurs optimales résultantes k_h par k_0 et n_0 respectivement, de sorte que l'accroissement optimal de l'échantillon dans la strate h soit $n_h - n_0$. On peut montrer que la minimisation de la taille totale d'échantillon sous toutes les contraintes $VR({}^d\hat{\bar{Y}}) \leq {}^dVR_0$, $d = 1, \dots, D$, $VR({}^d\bar{Y}) \leq {}^dVR_0$, et $0 < \tilde{f}_h \leq 1$, $h = 1, \dots, L$ mène à la même solution optimale, $n_0 = (n_1^0, \dots, n_L^0)^T$. Cependant, les exigences de fiabilité au niveau du domaine peuvent souvent être spécifiées après avoir déterminé n_0 .

5. Résultats empiriques

À la présente section, nous étudions les propriétés relatives des diverses méthodes de répartition de l'échantillon en nous servant de données tirées de l'EMCD. Aux sections 5.1 et 5.2, nous présentons nos résultats pour les estimateurs directs et les estimateurs composites des moyennes de strate, respectivement. À la section 5.3, nous donnons les résultats pour les moyennes de domaine.

5.1 Moyennes de strate : estimateurs directs

Pour l'étude empirique, nous avons utilisé un sous-ensemble des valeurs obtenues pour la population de l'EMCD limitée aux établissements uniques. Les tailles de strate, N_h , les moyennes de population de strate, \bar{Y}_h , les écarts-types de strate, S_h , et les CV de strate, $C_h = S_h / \bar{Y}_h$, sont données au tableau 1 pour les dix provinces du Canada (traitées comme des strates). Pour la répartition NLP, les tolérances que nous avons choisies pour les CV étaient $CV_{0h} = 15\%$ pour les moyennes de strate \bar{Y}_h et $CV_0 = 6\%$ pour la moyenne d'échantillon pondérée \bar{Y}_{st} désignée Canada (CA).

La répartition NLP satisfaisait aux tolérances de CV spécifiées à donné une taille globale minimale d'échantillon $n^0 = 3446$. Le tableau 2 montre la répartition de l'échantillon n_h^0 et les coefficients de variation $CV(\bar{Y}_h)$ et $CV(\bar{Y}_{st})$ associés pour la répartition NLP. Son examen montre que la répartition NLP respecte la tolérance spécifiée $CV_0 = 6\%$, donne des CV plus petits que la tolérance

d'ingéralité sont utilisés dans (2.6) et (2.7) parce que les CV résultants pour certaines strates h et/ou pour l'agrégat peuvent être plus petits que les tolérances spécifiées (Cochran 1977, page 122).

En posant que $k_h = f_h^{-1}$, l'expression (2.5) devient une fonction convexe séparable des variables k_h .

$$\tilde{g}(\mathbf{k}) = \sum_{h=1}^L N_h k_h^{-1}. \quad (2.9)$$

Nous spécifions à nouveau les contraintes (2.6) et (2.7) en fonction des variances relatives afin qu'elles soient linéaires en les variables k_h . La variance relative (VR) de \underline{y}_h est le carré de son CV,

$$VR(\underline{y}_h) = \frac{k_h - 1}{k_h} C_h^2. \quad (2.10)$$

De même, la variance relative de \underline{y}_{sr} est le carré de son CV,

$$VR(\underline{y}_{sr}) = \underline{y}_{sr}^{-2} \sum_{h=1}^L W_h^2 \frac{N_h}{k_h - 1} S_h^2. \quad (2.11)$$

Nous avons utilisé la procédure NLP de SAS avec l'option Newton-Raphson pour trouver la valeur optimale de k_h qui minimiserait (2.9) sous les contraintes

$$VR(\underline{y}_h) \leq VR^{oh}, \quad h = 1, \dots, L, \quad (2.12)$$

$$VR(\underline{y}_{sr}) \leq VR^{os}, \quad (2.13)$$

$$k_h \geq 1, \quad h = 1, \dots, L. \quad (2.14)$$

VR oh = VR $_2^{oh}$ et VR $_0^{oh}$ = VR $_2^{oh}$. En exprimant les contraintes sous forme linéaire et la fonction d'objectif sous forme d'une fonction convexe séparable, nous arrivons plus rapidement à la convergence du problème de NLP reformulé. En désignant la solution NLP comme étant $\mathbf{k}_0 = (k_1^0, \dots, k_L^0)^T$, le vecteur correspondant de tailles optimales d'échantillon de strate est donné par $\mathbf{n}_0 = (n_1^0, \dots, n_L^0)^T$, où $n_0^h = N^h / k_h^0$. Nous pouvons modifier (2.14) pour nous assurer que $n_0^h \geq 2$ pour tout h , ce qui permet d'obtenir une estimation sans biais des variables.

La méthode NLP peut être étendue facilement à plusieurs variables y_1, \dots, y_P en spécifiant des tolérances pour les CV des moyennes de strate et de la moyenne de population estimées pour chaque variable ($P = 1, \dots, P$). Si le nombre de variables P n'est pas faible, la taille d'échantillon totale optimale résultante $n_0^h = \sum_h n_h^0$ peut devenir considérablement plus grande que n_0^h pour une seule variable. Huddleston, Claypool et Hocking (1970), Bethel (1989) et d'autres chercheurs ont étudié la NLP pour obtenir une

3. Répartition pour les estimateurs composites

Longford (2006) a étudié les estimateurs composites des moyennes de strate de la forme

$$\theta_h = \alpha_h \underline{y}_{sr}^h + (1 - \alpha_h) \underline{y}_h^h \quad (3.1)$$

où \underline{y}_{sr}^h est un estimateur synthétique; ici, nous prenons $\underline{y}_{sr}^h = \underline{y}_{sr}$. L'EQM de θ_h est

$$EQM(\hat{\theta}_h) = V(\hat{\theta}_h) + [B(\hat{\theta}_h)]^2 \\ = \alpha_h^2 \sum_{l=1}^L \frac{W_l^h}{S_l^2} + (1 - \alpha_h)^2 W^h S_h^2 \\ + 2\alpha_h(1 - \alpha_h) \frac{W_h^h}{S_h^2} + \alpha_h^2 (\underline{y}_h^h - \underline{Y})^2$$

+ termes ne dépendant pas des n_h . (3.2)

Longford (2006) a montré que, dans (3.1), le coefficient optimal α_h^* qui minimise (3.2) est approximativement égal à $\alpha_h^* = S_h^h / (S_h^h + n_h \Delta_h^{-1})$, où $\Delta_h = \underline{y}_h^h - \underline{Y}$. Il a ensuite remplacé Δ_h^h dans α_h^* par sa moyenne sur les strates, désignée par $\bar{\omega}_h^2 = L^{-1} \sum_h (\underline{y}_h^h - \underline{Y})^2$, ce qui donne $\alpha_h^* \approx (1 + n_h \bar{\omega}_h^2)^{-1}$, où $\bar{\omega}_h^2 = \sigma_B^2 / S_h^h$. L'EQM résultante de $\hat{\theta}_h$ est approximée par

$$EQM(\hat{\theta}_h) \approx \frac{\sigma_B^2}{2} \frac{1 + n_h \bar{\omega}_h^2}{\sigma_B^2}. \quad (3.3)$$

La répartition de Longford s'obtient en minimisant la fonction d'objectif

$$\sum_{h=1}^L P_h EQM(\hat{\theta}_h) + (GP^+) V(\underline{y}_h) \quad (3.4)$$

par rapport aux n_h . La solution résultante satisfait à

$$\frac{P_h \sigma_B^2 \bar{\omega}_h^2}{2} + (GP^+) W_h^h S_h^2 = \text{const}, \quad h = 1, \dots, L. \quad (3.5)$$

Longford a utilisé une méthode itérative pour obtenir la solution de (3.5) puisqu'il n'y a pas de solution analytique. Notre procédure NLP minimise $g(\mathbf{f})$ donnée par (2.5) sous les contraintes

$$REQM(\hat{\theta}_h) \leq REQM^{oh}, \quad h = 1, \dots, L, \quad VR(\underline{y}_{sr}) \leq VR^{os} \quad (3.6)$$

dans l'étude numérique fondée sur des données provenant de l'Enquête mensuelle sur le commerce de détail (EMCD) de la Statistique Canada présentée à la section 5.

2.2 Répartition de Longford

La méthode de Longford (2006) vise à contrôler simultanément la fiabilité des estimations des moyennes de strate \bar{y}_h et de la moyenne de population \bar{y}_{st} en minimisant la fonction d'objectif

$$(2.3) \quad \sum_{h=1}^H P_h^+ V(\bar{y}_h) + (GP^+)(V(\bar{y}_{st}))$$

par rapport aux tailles d'échantillon de strate n_h sous la contrainte $\sum_{h=1}^H n_h = n$, où $P_h^+ = \sum_{h=1}^H P_h$. Dans (2.3), le premier terme spécifie l'importance relative, P_h , de chaque strate h , tandis que le deuxième accorde une importance relative à \bar{y}_{st} , au moyen du poids G . Longford (2006) a supposé que $P_h = N_h^q$ pour une certaine constante q ($0 \leq q \leq 2$). Le terme P_h dans (2.3) neutralise l'effet des tailles P_h et du nombre de strates sur le poids G .

Sous échantillonnage aléatoire simple stratifié, la répartition de l'échantillon qui minimise (2.3) est

$$(2.4) \quad n_L^h = n \frac{S_h \sqrt{f_h^*} \sum_{h=1}^H S_h \sqrt{f_h^*}}{S_h \sqrt{f_h^*}}, \quad h = 1, \dots, L$$

où $P_h^+ = P_h + GP^+ W_h^2$. Si $q = 2$, l'expression (2.4) ne dépend pas de la valeur de G et elle se réduit à la répartition de Neyman, n_N^h donnée par (1.2).

2.3 Répartition par programmation non linéaire (NLP)

Nous passons maintenant à la méthode NLP de détermination des tailles d'échantillon de strate n_h sous la contrainte des exigences de fiabilité spécifiées pour les moyennes d'échantillon de strate ainsi que pour la moyenne de population estimée. En posant que $\mathbf{f} = (f_1, \dots, f_L)^T$ avec $f_h = n_h / N_h$, nous minimisons la taille totale d'échantillon

$$(2.5) \quad g(\mathbf{f}) = \sum_{h=1}^H f_h N_h$$

par rapport à \mathbf{f} sous les contraintes

$$(2.6) \quad CV(\bar{y}_h) \leq CV_{0h}, \quad h = 1, \dots, L$$

$$(2.7) \quad CV(\bar{y}_{st}) \leq CV_0$$

$$(2.8) \quad 0 < f_h \leq 1, \quad h = 1, \dots, L$$

où CV_{0h} et CV_0 sont les tolérances spécifiées pour les CV de la moyenne d'échantillon de strate \bar{y}_h et de la moyenne de population estimée \bar{y}_{st} , respectivement. Des signes

des moyennes de strate \bar{y}_h peuvent être utilisés.

L'objectif principal du présent article est de proposer une méthode de répartition « optimale », fondée sur la programmation non linéaire (NLP, pour *non-linear programming*) (voir la section 2.3). La méthode consiste à minimiser la

taille totale d'échantillon $\sum_{h=1}^H n_h$ sous la contrainte de tolérances spécifiées pour les CV des moyennes d'échantillon de strate \bar{y}_h et de la moyenne de population estimée \bar{y}_{st} . Le cas des estimateurs indirects (composés) des moyennes de strate est étudié à la section 3. À la section 4,

nous examinons la répartition optimale de l'échantillon entre les strates quand des exigences de fiabilité pour des domaines recouvrant les strates sont également spécifiées. La méthode proposée s'étend facilement à plusieurs variables, mais pour simplifier, nous omettons les détails. À la section 5, en partant de la taille totale d'échantillon optimale obtenue par NLP, nous procédons à une étude numérique des propriétés des méthodes de Costa et coll. et de Longford en ce qui concerne le respect des exigences de fiabilité.

2. Répartition pour les estimateurs directs

À la présente section, nous considérons les estimateurs directs, \bar{y}_h , des moyennes de population de strate sous échantillonnage aléatoire simple stratifié. Le cas des estimateurs indirects des moyennes de strate est étudié à la section 3. Les estimateurs indirects sont utilisés lorsque les tailles d'échantillon de strate n_h sont faibles.

2.1 Répartition de Costa et coll.

La répartition de l'échantillon de Costa et coll. (2004) est

$$(2.1) \quad n_h^C = k(nW_h) + (1 - k)(n/L)$$

pour une constante spécifiée k ($0 \leq k \leq 1$). Cette répartition se réduit à l'équité répartition quand $k = 0$ et à la répartition proportionnelle quand $k = 1$. La formule (2.1) doit être modifiée quand $n/L > 1$ pour une strate h dans un ensemble de strates A . La répartition modifiée est

$$(2.2) \quad n_h^C = k(nW_h) + (1 - k)n_0^h$$

où $n_0^h = N_h$ si $h \in A$ et $n_0^h = (n - \sum_{h \in A} N_h) / (L - m)$, autrement où m est le nombre de strates dans l'ensemble A . Notons que, si $k = 0$, (2.2) donne l'équité répartition modifiée. Nous examinons différents choix de la constante k

À propos de la répartition de l'échantillon pour une estimation sur domaine efficace

G. Hussain Choudhry, J.N.K. Rao et Michael A. Hidiroglou¹

Résumé

Les questions concernant la répartition de l'échantillon sont étudiées dans le contexte de l'estimation des moyennes de sous-population (strate ou domaine), ainsi que de la moyenne de population agrégée sous échantillonnage aléatoire simple stratifié. Une méthode de programmation non linéaire est utilisée pour obtenir la répartition « optimale » de l'échantillon entre les strates qui minimise la taille totale d'échantillon sous la contrainte des tolérances spécifiques pour les coefficients de variation des estimateurs des moyennes de strate et de la moyenne de population. La taille totale d'échantillon résultante est alors utilisée pour déterminer les répartitions de l'échantillon par les méthodes de Costa, Satorra et Ventura (2004) s'appuyant sur une répartition intermédiaire ou de compromis et de Longford (2006) fondée sur des « priorités inférentielles » spécifiques. En outre, nous étudions la répartition de l'échantillon entre les strates quand sont également spécifiées des exigences de fiabilité pour des domaines qui recoupent les strates. Les propriétés des trois méthodes sont étudiées au moyen de données provenant de l'Enquête mensuelle sur le commerce de détail (EMCD) menée par Statistique Canada auprès d'établissements uniques.

Mots clés : Estimateurs composites ; répartition intermédiaire ; estimateurs directs ; moyennes de domaine ; programmation non linéaire.

1. Introduction

L'usage de l'échantillonnage simple stratifié est très répandu dans les enquêtes auprès des entreprises et d'autres enquêtes auprès des établissements employant des listes comme base de sondage. La moyenne de population $\bar{Y} = \sum_h W_h \bar{Y}_h$ est estimée par la moyenne d'échantillon pondérée $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, où $W_h = N_h/N$ est la taille relative de la strate h ($h = 1, \dots, L$), et \bar{Y}_h et \bar{y}_h sont les moyennes de la population et d'échantillon, respectivement, de la strate. La méthode de bien connue de Neyman en vue de répartir l'échantillon entre les strates est optimale pour l'estimation de la moyenne de population si l'on considère la minimisation de la variance de \bar{y}_{st} sous la contrainte que $\sum_h n_h = n$, où n est fixe, ou la minimisation de $\sum_h n_h$ sous la contrainte que la variance de \bar{y}_{st} est fixe, où n_h est la taille d'échantillon de la strate. Toutefois, la répartition de Neyman peut donner lieu à de grands coefficients de variation (CV) des moyennes \bar{y}_h . Par ailleurs, l'équité de l'échantillon, $n_h = n/L$, est efficace pour l'estimation des moyennes de strate, mais peut produire de beaucoup plus grands CV de l'estimateur \bar{y}_{st} que la méthode de Neyman.

$$n_h = n = \sum_{h=1}^L C_h X_h^h, \quad h = 1, \dots, L \quad (1.1)$$

telle est donnée par

$C_h = S_h/\bar{Y}_h$ est le CV de la strate, la répartition exponentielle de Neyman et l'équité de l'échantillon. En posant que la répartition de Neyman et l'équité de l'échantillon sont égales, on propose une « répartition exponentielle » (power allocation) comme compromis entre la

$$n_N = n = \sum_{h=1}^L N_h S_h^h, \quad h = 1, \dots, L \quad (1.2)$$

Neyman

où X_h est une mesure de taille ou d'importance de la strate h , et q est une constante de mise au point. La répartition exponentielle (1.1) s'obtient en minimisant $\sum_h \{X_h^q (CV(\bar{y}_h))^2\}$ sous la contrainte $\sum_h n_h = n$, où $CV(\bar{y}_h)$ est le CV de la moyenne d'échantillon de la strate h . Le choix $q = 1$ et $X_h = N_h \bar{Y}_h$ dans (1.1) mène à la répartition de

indirecte dont les valeurs de population sont connues. Costa et coll. (2004) ont proposé une répartition intermédiaire fondée sur une combinaison convexe d'une répartition proportionnelle, $n_h = n/L$, Longford (2006) a étudié systématiquement la répartition sous échantillonnage aléatoire simple stratifié en introduisant des « priorités

1. G.H. Choudhry, Division de la recherche et de l'innovation en statistique, Université du Québec à Montréal, 100, rue Saint-Jacques, Montréal, Québec H2Y 1A7, Canada. Courriel : gchoudhry@uqam.ca ; J.N.K. Rao, École de mathématiques et de statistique, Université du Québec à Trois-Rivières, 333, rue des Sciences, Trois-Rivières, Québec G9A 5H6, Canada. Courriel : jrao@math.usherbrooke.ca ; M.A. Hidiroglou, Division de la recherche et de l'innovation en statistique, Statistique Canada, Courriel : mike.hidiroglou@statcan.gc.ca.

Bibliographie

- Balakrishnan, V.K. (1997). *Graph theory*. New York : McGraw Hill, Inc.
- Beizer, B. (1995). *Black-box testing*. New York : John Wiley & Sons, Inc.
- Berge, C. (1976). *Graphs and hypergraphs*. New York : North-Holland Publishing Company - Amsterdam, London and American Elsevier Publishing Company, Inc.
- Bethlehem, J., et Hundepool, A. (2004). TADEQ: A tool for the documentation and analysis of electronic questionnaires. *Journal of Official Statistics*, 20, 233-264.
- Centers for Medicare and Medicaid Services (2010). *Medicare Current Beneficiary Survey: Overview*. Août 2002. Adresse URL : https://www.cms.gov/LimitedDataSets/11_MCBS.asp.
- Chatrand, G. (1985). *Introductory Graph Theory*. New York : Dover Publications, Inc., Mineola.
- Cochran, W.G. (1977). *Sampling Techniques*, 3^{ème} Edition. New York : John Wiley & Sons, Inc.
- Cohen, J. (1997). *Design and methods of the Medical Expenditure Panel Survey Household Component*. Rockville (MD) : Agency for Health Care Policy and Research. MEPS Methodology Report No. 1. AHCPR Pub. No. 97-0026.
- Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L. et O'Reilly, J.M. (1988). *Computer Assisted Information Collection*. New York : John Wiley & Sons, Inc.
- Gibbons, A. (1985). *Algorithmic Graph Theory*. Cambridge University Press.
- Harary, F., et Palmer, E. (1973). *Graphical Enumeration*. New York : Academic Press.
- Hetzl, W. (1983). *The Complete Guide to Software Testing, QED*. Massachusetts : Information Sciences, Inc., Wellesley.
- Kaner, C., Falk, J. et Nguyen, H. (1999). *Testing Computer Software*, 2^{ème} Edition. New York : John Wiley & Sons, Inc.
- Medical Expenditure Panel Survey (MEPS) (2010). *Survey Instruments and Associated Documentation*. <http://www.meepsahrq.gov/meepsweb/>.
- Myers, G.J. (1979). *The Art of Software Testing*. New York : John Wiley & Sons, Inc.
- Pattou, R. (2006). *Software Testing*, 2^{ème} Edition. Sams Publishing, Inc.
- Poole, J. (1995). NISTIR 5737 – A Method to Determine a Basis Set of Paths to Perform Program Testing. National Institute of Standards and Technology, Gaithersburg, MD, Novembre, 1995.
- Statistique Canada (2010). *Enquête Canadienne sur les capacités fonctionnelles (ECCF): Questionnaire 2009*. <http://www.statcan.gc.ca/nanctives/ECCF/Questionnaire2009>.
- Statistics Netherlands (2002). *Blaise Developer's Guide*. Department of Statistical Informatics, Statistics Netherlands, Heerlen.
- US Bureau of Labor Statistics (2010). *Consumer Expenditure Surveys Quarterly Interview API Survey 2010*. United States Department of Labor, <http://www.bls.gov/cex/capi/2010/cecapihome.htm>.
- Watson, A., et McCabe, T. (1996). NIST Special Publication 500-235 Structured Testing: A Testing Methodology Using the Cyclomatic Complexity Metric. National Institute of Standards and Technology, Gaithersburg, MD, Septembre, 1996.

Les cellules indiquent le nombre total de liens entre nœuds successifs d'après la matrice des parcours de base ci-avant.
Parcours de base = 9 Nombre total de chemins = 1 504

Chemin	Appariement des nœuds									
	1 ^{er} au 2 ^e	2 ^e au 3 ^e	3 ^e au 4 ^e	4 ^e au 5 ^e	5 ^e au 6 ^e	6 ^e au 7 ^e	7 ^e au 8 ^e	8 ^e au 9 ^e	9 ^e au 10 ^e	10 ^e au 11 ^e
Parcours de base 1	2	2	3	4	2	2	2	2	2	2
Parcours de base 2	2	2	4	3	4	2	2	2	2	2
Parcours de base 3	2	4	3	5	3	2	2	2	2	2
Parcours de base 4	2	2	2	5	4	2	2	2	2	2
Parcours de base 5	2	4	4	2	4	2	2	2	2	2
Parcours de base 6	2	2	5	5	4	2	2	2	2	2
Parcours de base 7	2	4	4	3	4	2	2	2	2	2
Parcours de base 8	2	2	4	2	2	2	2	2	2	2
Parcours de base 9	2	4	2	2	2	2	2	2	2	2

Tableau 6
Nombre de liens et nombre de chemins pour chaque parcours de base

Les valeurs dans les cellules représentent les nœuds. Chaque ligne correspond à un parcours de base.

Chemin	Appariement des nœuds									
	1 ^{er} au 2 ^e	2 ^e au 3 ^e	3 ^e au 4 ^e	4 ^e au 5 ^e	5 ^e au 6 ^e	6 ^e au 7 ^e	7 ^e au 8 ^e	8 ^e au 9 ^e	9 ^e au 10 ^e	10 ^e au 11 ^e
Parcours de base 1	1	2	3	6	9	16	16	16	16	16
Parcours de base 2	1	2	4	6	9	16	16	16	16	16
Parcours de base 3	1	2	5	7	10	16	16	16	16	16
Parcours de base 4	1	2	4	7	10	16	16	16	16	16
Parcours de base 5	1	2	5	8	12	16	16	16	16	16
Parcours de base 6	1	2	5	7	11	16	16	16	16	16
Parcours de base 7	1	2	5	8	13	16	16	16	16	16
Parcours de base 8	1	2	5	8	14	16	16	16	16	16
Parcours de base 9	1	2	5	8	15	16	16	16	16	16

Tableau 5
Matrice des parcours de base

Les cellules indiquent le nombre total de liens entre nœuds successifs à l'intérieur d'un parcours.

Chemin	Appariement des nœuds									
	1 ^{er} au 2 ^e	2 ^e au 3 ^e	3 ^e au 4 ^e	4 ^e au 5 ^e	5 ^e au 6 ^e	6 ^e au 7 ^e	7 ^e au 8 ^e	8 ^e au 9 ^e	9 ^e au 10 ^e	10 ^e au 11 ^e
Parcours 1	2	2	3	4	2	2	2	2	2	2
Parcours 2	2	2	4	4	2	2	2	2	2	2
Parcours 3	2	4	3	5	3	2	2	2	2	2
Parcours 4	2	2	2	5	3	2	2	2	2	2
Parcours 5	2	4	4	2	4	2	2	2	2	2
Parcours 6	2	2	5	5	4	2	2	2	2	2
Parcours 7	2	4	3	4	4	2	2	2	2	2
Parcours 8	2	2	4	2	2	2	2	2	2	2
Parcours 9	2	4	2	2	2	2	2	2	2	2
Parcours 10	2	4	2	2	2	2	2	2	2	2

Tableau 4
Nombre de liens et nombre de chemins pour chaque parcours

Les valeurs dans les cellules représentent les nœuds. Chaque ligne correspond à un parcours.
[Remarque : Dans cet exemple, les parcours comptent tous six nœuds. Toutefois, de manière générale, les parcours n'auront pas tous le même nombre de nœuds.]

Chemin	Appariement des nœuds									
	1 ^{er} au 2 ^e	2 ^e au 3 ^e	3 ^e au 4 ^e	4 ^e au 5 ^e	5 ^e au 6 ^e	6 ^e au 7 ^e	7 ^e au 8 ^e	8 ^e au 9 ^e	9 ^e au 10 ^e	10 ^e au 11 ^e
Parcours 1	1	2	3	6	9	16	16	16	16	16
Parcours 2	1	2	4	6	9	16	16	16	16	16
Parcours 3	1	2	5	7	10	16	16	16	16	16
Parcours 4	1	2	4	7	10	16	16	16	16	16
Parcours 5	1	2	5	8	12	16	16	16	16	16
Parcours 6	1	2	5	7	11	16	16	16	16	16
Parcours 7	1	2	5	8	13	16	16	16	16	16
Parcours 8	1	2	5	8	14	16	16	16	16	16
Parcours 9	1	2	5	8	15	16	16	16	16	16
Parcours 10	1	2	5	8	16	16	16	16	16	16

Tableau 3
Matrice des parcours

erreurs.

Il convient de souligner l'appui de Westat Incorporated, de Rockville (Maryland) à l'égard d'une partie importante des travaux ayant abouti à la rédaction de ce document. Mentionnons également la collaboration des réviseurs et des rédacteurs de *Techniques d'enquête* ainsi que les améliorations qu'ils ont apportées au document.

Exemple de génération d'une base

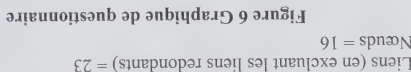


Figure 6 Graphique de questionnaire

Matrice des liens	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
Nœud	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Tableau 2

Chaque cellule contient une valeur relative au nombre de liens entre les nœuds figurant dans les colonnes et ceux figurant dans les lignes.

qu'ils contiennent des erreurs d'enchaînement. Cela tient à la nature des travaux de programmation entourant la création d'un instrument d'IAO. Les options de réponse sont habituellement « groupées », en ce sens que les options qui font passer à la même question suivante seront probablement toutes soit mal orientées, soit bien orientées. C'est pourquoi soumettre une base à des essais complets constitue une méthode efficace de limiter les erreurs les plus susceptibles de donner lieu à la perte de données.

Cela dit, il se peut que des essais non exhaustifs constituent la seule stratégie raisonnable s'il existe des contraintes au chapitre du temps ou des efforts requis et si la base compte un grand nombre de parcours. Malgré le fait que n'importe quelle partie d'un instrument qui n'a pas fait l'objet d'essais peut contenir une erreur, toute portion de parcours dans la base peut constituer un essai non biaisé. Par conséquent, le pourcentage de parcours à inclure dans les essais dépendra probablement de facteurs propres à la situation conceptuelle particulière. Par exemple, un instrument peut compter des modules ayant été utilisés antérieurement ou qui n'ont fait l'objet que de modifications mineures depuis leur utilisation précédente. Il ne sera pas nécessaire de soumettre ces modules à des essais aussi poussés que s'il s'agit de modules entièrement nouveaux. En règle générale, un échantillon de cas types de la taille minimale requise devrait englober chaque section de l'instrument faisant partie d'un ou de plusieurs parcours, et les parcours inclus devraient couvrir toutes les connexions entre sections.

7. Discussion et conclusion

Une approche de développement de logiciels fondée sur la théorie des graphes présente deux avantages par rapport aux approches conventionnelles. D'abord, on peut ainsi disposer d'un système de documentation capable de simuler le déroulement d'une interview assistée par ordinateur. Cette caractéristique sera utile pour vérifier les enchaînements et aidera les évaluateurs à comparer directement le déroulement de l'instrument et les spécifications conceptuelles. L'autre grand avantage tient à la sélection de cas aux fins des essais. L'utilisation de la base d'un questionnaire rend possible la spécification d'un univers de cas types couvrant toutes les paires de nœuds à partir du nombre minimal de parcours requis. L'échantillonnage probabiliste effectué à partir de cet univers prévient tout biais dans les procédures d'essai.

Dans la pratique, on peut obtenir le premier de ces deux avantages en structurant la base de données d'après un système de spécifications qui fera en sorte que l'on dispose d'un tableau des nœuds et d'un tableau des liens. Si le tableau des liens précise le nœud prédécesseur et le

nœud successeur, les recherches dans les tableaux permettent de procéder à la vérification des enchaînements et à des simulations. L'autre avantage peut être obtenu au moyen d'un algorithme servant à déterminer une base. Ainsi que l'a mentionné Poole (1995), l'un des principales choses à faire lorsque l'on prévoit faire l'essai d'un logiciel consiste à déterminer quels cas types utiliser. Poole a présenté un algorithme qui accomplit cette tâche à partir du graphique de déroulement du programme. L'utilisation d'un tel graphique à cette fin est appropriée dans la mesure où le programme n'est pas de trop grande taille. Dans le cas de programmes compliqués et de grande envergure, les diagrammes de déroulement deviennent trop complexes. Il en va de même avec les questionnaires très étendus et compliqués (Bethlehem et Hundepool 2004). On trouvera en annexe le résultat obtenu au moyen d'un algorithme qui, à partir d'un graphique de questionnaire servant d'exemple, produit une base, fait le compte des chemins et détermine les parcours de base. (Il est possible d'obtenir cet algorithme en communiquant avec l'auteur (sdelliot12@verizon.net). L'algorithme en question ne traite pas les structures en boucle inhérentes aux contrôles ou aux fonctions de retour en arrière. Ces structures peuvent faire l'objet d'essais en dehors du graphique du questionnaire. Un algorithme traitant les boucles est en cours d'élaboration.)

Une approche reposant sur la théorie des graphes est en outre utile parce qu'elle permet de recourir à différentes mesures descriptives des questionnaires, comme le nombre de chemins, le nombre de parcours, la complexité cyclomatique (soit la mesure de la complexité du code d'un logiciel (se reporter à Hetzel 1984, McCabe 1976, Watson et McCabe 1996) ; elle est égale au nombre de parcours qui constituent la base du graphique. Dans le cas d'un graphique cyclomatique $(CC) = L - N + 2$, où L est le nombre de liens et N , le nombre de nœuds) et plusieurs types de matrices descriptives (se reporter à l'annexe). Les améliorations qui seront apportées aux approches fondées sur la théorie des graphes dans l'avenir prendront probablement la forme, par exemple, 1) de taxonomies relatives aux éléments, aux liens et aux erreurs, 2) de tableaux secondaires dans la base de données des spécifications concernant les attributs propres aux différents types de nœuds et de liens, 3) de plans d'échantillonnage sophistiqués pour la sélection de cas types et 4) du recours à l'échantillonnage des chemins par choix raisonné.

Les taxonomies favoriseront la spécification de types particuliers d'éléments des instruments ainsi que l'incorporation de tableaux secondaires au système de documentation. Un type particulier d'éléments comprendra par exemple des éléments comportant une fonction de randomisation.

Orphan Report

General Cancer Knowledge

These instruments were developed in the context of the...
...that there is a need for a...
...that there is a need for a...
...that there is a need for a...

Sequence	Instrument	Component	Count
1	Instrument 1	CK-1	0
14	Instrument 1	CK-8	1
12	Instrument 1	CK-7	1
16	Instrument 1	CK-9	1
21	Instrument 1	CK-13	1
23	Instrument 1	CK-14	1
25	Instrument 1	CK-15a	1
36	Instrument 1	CK-15k	1
49	Instrument 1	CK-18	3
45	Instrument 1	CK-16b	3
46	Instrument 1	CK-16c	3

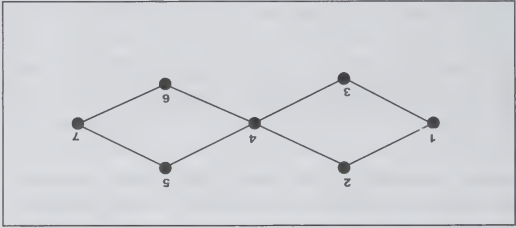
Figure 4 Rapport sur les orphelins

faire l'essai de tous les chemins serait fastidieux, du fait du grand nombre de chemins, mais aussi redondant, étant donné l'existence de nombreux chemins similaires. Le concepteur de l'essai aura donc pour tâche de choisir un sous-ensemble de chemins afin de maximiser la couverture et de limiter le plus possible la redondance. Il est possible d'y arriver en utilisant au départ le sous-ensemble **PB** à titre de première étape de l'échantillonnage à partir de l'univers de chemins. Utiliser **PB** ainsi équivaut à amorcer le processus d'échantillonnage avec un échantillon par choix raisonné (Cochran 1977, page 10). On peut aussi concevoir des éléments dans le but d'éliminer la redondance. L'univers ainsi constitué a une couverture complète et contient le plus petit ensemble de cas requis pour inclure toutes les paires de nœuds connectés. L'étape suivante pourrait consister à choisir un ou plusieurs chemins à partir de chacun des parcours faisant partie du sous-ensemble **PB**, ce que l'on peut faire de plusieurs manières, notamment en envisageant chaque parcours comme une grappe de cas types puis en procédant à un échantillonnage probabiliste à partir de chaque grappe. Ou encore, il serait possible de choisir un chemin dans chaque grappe en sélectionnant au hasard un lien parallèle pour chaque nœud.

À partir du moment où l'on accepte la notion d'essai portant sur la base, il faut déterminer quelle portion de la base doit faire l'objet des essais. Si les essais portent sur tous les parcours du sous-ensemble **PB**, les seuls éléments redondants. Ces liens peuvent comporter des fautes d'orthographe ou des erreurs de formatage, mais il est peu probable

Que l'on élimine n'importe quel de ces quatre parcours, les trois autres incluront chaque paire de nœuds connectés ; dès lors, n'importe quel trio de parcours constituera un ensemble de parcours de base (**PB**). Ainsi, si le parcours 1 est éliminé, chaque paire de nœud continuera d'être incluse dans les parcours 2, 3 et 4. Par contre, si l'on élimine les parcours 1 et 2, les paires de nœud 1 - 2 et 2 - 4 seraient exclus. L'ensemble de deux parcours ne suffirait donc pas à couvrir l'ensemble des séquences indépendantes de nœuds dans le graphique.

Figure 5 Représentation des parcours et de l'ensemble de parcours de base



chaque paire de nœuds que compte le parcours. Ainsi, la figure 1, le nombre de chemins pour chaque parcours est le suivant :

Parcours 1 - $2 \times 3 \times 2 \times 2 \times 2 \times 3 = 288$
Parcours 2 - $2 \times 3 \times 2 \times 2 \times 2 \times 3 = 288$
Parcours 3 - $2 \times 2 \times 2 \times 2 \times 3 = 96$

Le nombre total de chemins correspond à la somme des chemins associés à tous les parcours (dans notre exemple, $288 + 288 + 96 = 672$). On peut effectuer ce calcul au moyen de la formule suivante :

$$\text{Chemins} = \sum_i P_{NP_i} \text{liens}_i$$

où i représente le i^{e} parcours, P est le nombre total de parcours, f est le f^{e} ensemble de liens à l'intérieur d'un parcours donné, NP_i est le nombre de paires de nœuds liés entre eux à l'intérieur d'un parcours donné et liens est le nombre de liens unissant une paire de nœuds.

Si un protocole d'essai repose sur un échantillon de chemins, la base du graphique comportera un ensemble à la fois minimal et exhaustif, ou un univers, de cas types. Dans ce contexte, le terme « base » a un sens analogue à celui applicable en géométrie. En effet, la base d'un espace géométrique est composée d'un ensemble de vecteurs qui quel point à l'intérieur de l'espace. De même, la base d'un graphique est un ensemble de parcours qui suffit à englober toutes les paires de nœuds prédécesseurs-successeurs. Cela signifie que la base comprend tous les nœuds ainsi qu'au moins un des liens reliant chaque paire de nœuds connectés entre eux. Une base constituera un sous-ensemble de l'ensemble des parcours possibles. Tout questionnaire comporte un ensemble de parcours (P) dont chaque membre satisfait à la définition des parcours énoncée précédemment (séquence unique de nœuds). À l'intérieur de cet ensemble, il y a un sous-ensemble dont la caractéristique est que chaque parcours qui en fait partie contient au moins une paire de nœuds connectés qui n'est contenue dans aucun autre parcours du sous-ensemble. Il s'agit du sous-ensemble de « parcours de base » (PB).

Pour mieux comprendre la différence entre les parcours faisant partie du sous-ensemble PB et ceux du sous-ensemble qui constitue son complément (soit $P - PB$), examinons le graphique présenté à la figure 5. L'ensemble des parcours (P) du graphique est :

Parcours 1 - 1, 2, 4, 5, 7
Parcours 2 - 1, 2, 4, 6, 7
Parcours 3 - 1, 3, 4, 5, 7
Parcours 4 - 1, 3, 4, 6, 7

6. Approche d'essai fondée sur la théorie des graphes

Il existe plusieurs moyens de définir un univers d'éléments d'essai. On peut par exemple utiliser les cas types, dont il a déjà été question, chaque cas type étant une interview fictive. Il pourrait aussi s'agir de questions d'enquête ou d'options de réponse, ou encore d'une combinaison quelconque des unes et des autres. Nos commentaires porteront uniquement sur les cas types, de sorte qu'il est utile de donner une définition précise des cas types et de deux termes étroitement liés, soit le « parcours » et le « chemin ».

Un parcours se compose d'un ensemble unique et ordonné de nœuds allant du début à la fin de l'instrument. Chaque nœud faisant partie d'un parcours donné, exception faite d'un nœud de début ou de fin, est lié à la fois à un prédécesseur et à un successeur (cette définition concorde avec les travaux de Bethlehem et Hundepool 2004). Un parcours unique existe chaque fois qu'un élément compte plus d'un successeur. À la figure 1, il existe de multiples successeurs dans le cas des éléments 2 et 4. Ces deux points d'embranchement engendrent trois parcours :

Parcours 1 - 1, 2, 3, 5, 7, 9, 11, 12
Parcours 2 - 1, 2, 3, 5, 8, 9, 11, 12
Parcours 3 - 1, 2, 4, 6, 10, 11, 12

Un chemin constitue pour sa part un ensemble unique de nœuds et de liens alternés à partir du nœud du début jusqu'au nœud de la fin. Tout comme un parcours, un chemin doit satisfaire aux conditions de connectivité et d'orientation. En théorie des graphes, ce terme est synonyme de cas type ou encore de scénario d'essai dans le domaine des essais de logiciels. Étant donné que le chemin englobe le lien unissant une paire de nœuds, le nombre de chemins à l'intérieur d'un graphique sera supérieur ou égal au nombre de parcours. Le nombre de chemins que contient un parcours donné est égal au produit du nombre de liens entre

Simulate :

General Cancer Knowledge

Instrument 1

Component Name CK-3

Sequence No: 3

Description:

Type: Field

▼

CK-3. And which of the four remaining illnesses causes the second greatest number of deaths? [NOTE: Display the four response alternatives not selected in the previous question.]

Instrument Component

Instrument 1 CK-4a

Go To Origin

View Responses

Close

Figure 3 Ecran de simulation

5. Essai

Le processus d'essai d'un instrument d'enquête assisté par ordinateur consiste à vérifier que l'exécution de l'instrument est conforme aux spécifications conceptuelles. Plusieurs approches ont été utilisées à cette fin. L'une d'entre elles consiste à tester d'abord les éléments de base du système, puis de passer à des ensembles de plus en plus grands et intégrés d'éléments (essai ascendant). On parle d'« essai d'unités » dans le cas des essais portant sur les éléments de base (Beizer 1995, page 5). Une fois que chaque élément de base a fait l'objet d'un essai exclusif, les éléments sont assemblés et les essais portent sur leurs interactions. C'est ce que l'on appelle l'« essai d'intégration » (Hetzel 1984, page 11), dont l'étape finale est l'essai du système dans son ensemble, en conformité avec l'utilisation qui en serait faite dans un environnement de production (Myers 1979, page 110).

D'autres approches et d'autres termes ont été employés en ce qui touche les procédures d'essai, par exemple les méthodes de la boîte noire et de la boîte blanche ainsi que les essais de régression. Dans la méthode de la boîte noire, c'est-à-dire que son fonctionnement interne ne peut être observé. Les intrants et les extrants sont les seuls aspects observables du fonctionnement du programme (Beizer 1995, page 8). Dans la méthode de la boîte blanche, on se fonde sur la connaissance du code du programme pour décider de la manière de procéder aux essais et des cas qui feront l'objet des essais (Patton 2006, page 55). Par exemple, un programmeur peut mener un ensemble d'essais selon la méthode de la boîte blanche, de manière à exécuter chaque ligne de code (« couverture du code ») ou chaque point d'embranchement (« couverture des points d'embranchement »). Les essais de régression ont pour objet de garantir l'intégrité du code lorsque des ajouts ou des changements sont apportés à un programme opérationnel (Beizer 1995, page 235). Les essais de régression portent sur un ensemble de cas types choisis de manière à assurer l'exécution de chacun des principaux embranchements du programme. D'autres types d'essais sont utilisés dans le cadre de l'élaboration de logiciels (versions alpha et bêta, conviabilité), et de nombreuses sources fournissent une description plus complète des procédures d'essai (se reporter à Kaner, Falk et Nguyen 1999, page 277).

Dans toute procédure d'essai, l'une des principales préoccupations aura trait au biais d'essai. Cela survient lorsque certains éléments ou certaines fonctionnalités d'un instrument sont exclus des essais, ce qui peut arriver par exemple dans le cas de questions posées vers la fin d'une enquête ou faisant partie d'une section plus obscure. Ce biais est entièrement éliminé si l'on sélectionne un ensemble de cas types de manière à inclure l'ensemble des éléments de l'instrument, des liens entre ces éléments et des aspects fonctionnels. Cela dit, compte tenu de la longueur et de la complexité de certaines enquêtes, il n'est pas judicieux dans

vérifier la formulation et la présentation des questions ainsi que les options de réponse ; elle permet aussi de vérifier que l'instrument assure la transition vers la question appropriée au moment approprié. Les rapports d'erreurs ou de problèmes peuvent être intégrés directement au système de spécifications à titre d'attributs d'un élément de l'instrument.

Une autre méthode d'évaluation de l'intégrité d'un questionnaire consiste à déterminer les éléments « orphelins » de l'instrument. En effet, il peut arriver lors de la création ou de la modification d'un questionnaire qu'un élément devienne inaccessible – ce que l'on désignera au moyen du terme « orphelin ». Étant donné qu'il existe un tableau des liens (options de réponse et conditions), il est possible de lancer des recherches à partir de ce tableau afin de déterminer si une question donnée constitue le successeur d'un lien quelconque. La question qui ne succède à aucun lien est un « orphelin ». On trouvera à la figure 4 l'écran affichant la liste des éléments classés selon la fréquence à laquelle ils constituent des successeurs (ce que l'on a appelé le « rapport sur les orphelins »). On peut voir que la première question de l'enquête n'a pas de point d'origine, ce qui est normal puisque le premier élément ne peut avoir de prédécesseur. Tout autre élément dépourvu de tout point d'origine sera un orphelin. Le rapport sur les orphelins sert également à décrire les éléments de l'instrument. Par exemple, une question ou un élément comportant un grand nombre de points d'origine peut être la première question d'une section prévue dans les cas d'interruption prématurée. Une telle section sera accessible à partir de n'importe quelle autre section que comporte l'interview, de sorte qu'elle aura un grand nombre de prédécesseurs.

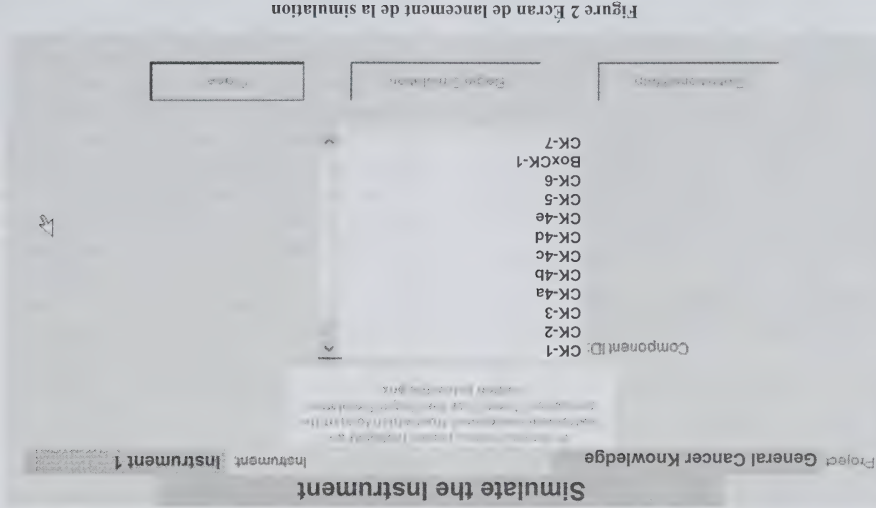


Figure 2 Écran de lancement de la simulation

écran. La figure 3 représente l'écran de simulation proprement dit. L'élément pertinent est affiché au centre de l'écran, ainsi que le texte de la question ou la condition applicable. La partie inférieure gauche de l'écran présente l'ensemble des éléments à partir desquels on peut avoir fait la transition vers l'élément affiché (prédécesseurs). On partera de « points d'origine » dans l'écran. La partie inférieure droite est réservée aux « points de destination », c'est-à-dire les éléments vers lesquels il est possible d'aller à partir de l'élément affiché (successeurs). Il est donc possible de cheminer dans un instrument à la fois dans l'une ou l'autre direction, un élément à la fois, en choisissant soit un point d'origine, soit un point de destination. Le questionnaire utilisé à titre d'exemple dans les figures 2 et 3 a trait aux connaissances générales sur le cancer, et la question exposée à la figure 4 compte un seul prédécesseur et un seul successeur. Ce sera le cas de la plupart des questions d'enquête ; si toutefois il existait de multiples prédécesseurs ou successeurs, l'écran en afficherait la liste.

Il est possible de simuler l'application d'un instrument d'enquête parce qu'un tableau distinct est utilisé pour les liens. Ce tableau peut se prêter à des recherches afin de trouver tous les prédécesseurs et successeurs de n'importe quel élément du questionnaire. Au cours de la phase de conception, cette caractéristique peut servir à s'assurer que chacune des sections et des questions est connectée comme il se doit et que tous les enchaînements sont exacts. Lors de la phase d'essai, la même caractéristique peut servir à effectuer une comparaison en parallèle d'un instrument et des spécifications sur lesquelles il repose. L'évaluateur peut recourir au système de spécifications pour simuler l'instrument sur un moniteur tout en exécutant l'instrument sur un second moniteur. Une telle comparaison peut servir à

Les caractéristiques d'un graphe de questionnaire peuvent être résumées comme suit :

1. Présence d'un nœud marquant le début et d'un autre marquant la fin.
2. Connectivité (chaques nœud est connecté aux nœuds de début et de fin).
3. Toutes les arêtes sont orientées.
4. Les paires de nœuds peuvent être connectées par des arêtes multiples ou parallèles.
5. Les nœuds peuvent figurer plus d'une fois à l'intérieur d'un même chemin.

À partir d'un ensemble de propriétés définitionnelles, il est possible d'établir différents descripteurs, dont le nombre de chemins et une base. Il est également possible de modéliser un système de documentation d'après la structure du graphe, ainsi qu'on le verra à la section suivante.

3. Systèmes de documentation et de spécifications des questionnaires d'enquête

Les systèmes de documentation de questionnaires entrent généralement dans deux catégories selon qu'il s'agit d'un document-texte ou d'une base de données relationnelle. Dans le cas d'un système qui consiste en un document-texte, l'information relative à une question de fond ou à une autre catégorie d'éléments de l'instrument sera le plus souvent présentée sous forme de section comportant le texte de la question, les options de réponse, les enchaînements et des instructions à l'intention des programmeurs. Le système de documentation n'a en soi aucune fonctionnalité, si ce n'est les fonctions de recherche et d'impression que comporte le logiciel de traitement de texte ayant servi à rédiger la documentation. Pour leur part, les systèmes reposant sur une base de données relationnelle sont habituellement organisés sous forme de tableau dont les lignes représentent les questions de l'enquête et les colonnes, les attributs des questions. Chaque enregistrement du tableau correspond à un n-uple des attributs des questions. Par exemple, les attributs d'une question peuvent comprendre l'intitulé, le numéro d'ordre, le texte de la question, les options de réponse, l'information sur les enchaînements et les notes techniques pertinentes. Le *Tool for the Analysis and Documentation of Electronic Questionnaires* (TADEQ) (Bethlehem et Hundepool 2004) est un système de spécifications correspondant à cette description. On peut aussi mentionner les systèmes conçus à Westat Inc. pour le *Medicare Current Beneficiary Survey* (MCBS), commandité par les US Centers for Medicare and Medicaid Services (*Medicare Current Beneficiary Survey: Overview* 2010), ou encore le *Medical Expenditure Panel Survey* (MEPS), commandité par le US Department of Health and Human

Services (*MEPS: Survey Instruments and Associated Documentation* 2010). Ces systèmes de base de données ont en commun une structure comportant un tableau primaire où chaque enregistrement représente une question.

Malgré les avantages découlant de la nature simple des systèmes conventionnels, un système de spécifications modélisé sous forme de graphe présentera de plus grandes capacités. Avant de décrire ces capacités et la structure sous-jacente requise, il convient de souligner qu'il existe beaucoup de façons d'élaborer une structure de données selon la théorie des graphes (on se reportera à ce sujet à Gibbons (1985, page 73), qui a décrit et catégorisé certaines de ces structures). Le système proposé ici consiste en une structure relationnelle (liste) comportant deux tableaux principaux. Le premier tableau représente les nœuds du graphe et le second, les arêtes. Dans le tableau des nœuds, chaque enregistrement (ligne) représente un élément donné de l'instrument (question d'enquête, contrôle, point de décision déterminant l'enchaînement). Dans le second tableau, chaque enregistrement représente une arête donnée (option de réponse ou condition particulière applicable à un point de décision). Chaque enregistrement faisant partie de l'un des tableaux comporte des attributs qui lui sont propres. Les différents attributs sont énoncés dans les colonnes du tableau. Dans le tableau des nœuds, chaque colonne représente un attribut particulier, par exemple l'identificateur et la catégorie de l'élément. Dans le tableau des arêtes, chaque arête (lignes) contient des identificateurs concernant le prédécesseur et le successeur d'une arête. Ainsi que cela est exposé à la section suivante, ces distinctions font en sorte qu'un tel système de documentation se prête à des utilisations que ne permet pas un système conventionnel.

4. Caractéristiques d'un système de spécifications fondé sur la théorie des graphes

L'utilisation de tableaux distincts pour les nœuds et pour les liens à titre d'éléments de base d'un système de spécifications présente plusieurs avantages, en particulier la possibilité de simuler une interview. Le concepteur ou l'évaluateur peut sélectionner les options de réponse de l'instrument au fil de l'enchaînement des éléments de ce dernier comme s'il faisait remplir l'instrument par un répondant. La figure 2 constitue un exemple d'écran de lancement de la simulation d'un instrument. L'élément qui marque le début de la simulation est choisi à partir de cet

Le présent document porte sur l'application de la théorie des graphes aux instruments de recherche et d'enquête. La section 2 propose la description d'un questionnaire sous forme de graphe et délimite les priorités particulières qui distinguent un graphe de questionnaire des autres types de graphes. La section 3 énonce les conséquences d'une représentation fondée sur la théorie des graphes de la structure des bases de données associées aux systèmes de documentation et de spécifications des enquêtes assistées par ordinateur. Les caractéristiques des structures de données selon la théorie des graphes sont commentées à la section 4. Les sections 5 et 6 portent sur les essais de logiciels et les conséquences associées à la théorie des graphes dans l'optique des essais. Un exposé est fait des raisons justifiant l'utilisation d'une « base » constituée d'un ensemble de cas types englobant toutes les paires de nœuds unis par des liens. Les parcours que comporte cette base constituent un ensemble complet de cas types aux fins de procéder à l'essai d'instruments.

2. Envisager le questionnaire comme un graphe

Il est possible de représenter un graphe de la façon suivante : $G = (V, E)$, où $V = \{v_1, v_2, v_3, \dots, v_n\}$ est un ensemble de nœuds ou de sommets, et $E = \{(v_i, v_j), (v_j, v_i), \dots\}$ est un ensemble de liens, ou de relations, entre paires de sommets. Dans le contexte de la théorie des graphes, on utilise le terme « arêtes » pour désigner les liens, qui sont représentés au moyen du symbole E (Chartrand 1985, page 27). Un graphe n'a pas à comporter d'autres caractéristiques particulières. Cela dit, un graphe auquel on incorpore des caractéristiques spéciales sera utile pour modéliser de nombreux phénomènes en sciences et en génie. À titre d'exemple, les graphes comportant des arêtes non orientées (c'est-à-dire où les deux nœuds liés entre eux peuvent l'un comme l'autre être le prédécesseur ou le successeur) peuvent servir à modéliser des circuits électriques à courant alternatif, tandis que les graphes dont les arêtes sont orientées pourront se prêter à la modélisation des problèmes entourant la conception des circuits de circulation. D'autres encore seront utilisés pour modéliser des réseaux dans des domaines comme l'informatique, les communications, la sociologie et la psychologie.

Dans le cas des questionnaires d'enquête, les nœuds du graphe représentent différents éléments de l'instrument d'enquête. Le plus souvent, il s'agit de questions de fond d'enquête ou de points de décision qui servent à déterminer l'enchaînement des questions. Les arêtes représentent pour leur part les options de réponse ou les résultats associés à un nœud. Elles représentent également le cheminement d'un nœud à l'autre, chaque arête comportant

un nœud prédécesseur et un nœud successeur uniques. Le graphe présenté à la figure 1 représente un instrument d'enquête simple comportant douze questions. Les cercles noirs (c'est-à-dire les nœuds) correspondent aux éléments de l'instrument, et les lignes qui relient ces cercles sont les arêtes unissant une question donnée à une autre question. Par exemple, le premier nœud peut représenter une question à laquelle deux réponses sont possibles, comme « oui » et « non ». Le deuxième nœud peut être une question avec cinq réponses possibles, les trois premières réponses étant liées à un lien avec le nœud 3 et les deux dernières, avec le nœud 4. Le graphe qui est utilisé pour représenter un questionnaire se voit attribuer différentes propriétés particulières. Ces propriétés définissent la nature logique du questionnaire. Bethchem et Hindepool (2004) ont exposé un certain nombre de ces propriétés. En premier lieu, un questionnaire comporte un nœud qui marque le début et un autre qui marque la fin. En deuxième lieu, tous les nœuds, autres que celui du début et celui de la fin, sont liés, c'est-à-dire que, pour chaque nœud du graphe, il y a au moins un chemin vers le nœud du début et au moins un vers le nœud de la fin. En troisième lieu, toutes les arêtes sont orientées. Cela veut dire que le chemin descendant de l'enchaînement d'un nœud à l'autre va toujours dans la même direction. En quatrième lieu, il peut exister plus d'une arête entre une paire donnée de nœuds. Dans de nombreux types de graphes, une paire de nœuds donnée sera reliée par une seule arête. Cette restriction ne s'applique toutefois pas dans le cas d'un graphe de questionnaire parce qu'un questionnaire comportera souvent plus d'une option de réponse déterminant l'enchaînement d'une question à une autre. Enfin, en cinquième lieu, il peut y avoir des structures en boucle, de sorte qu'un nœud pourra figurer plus d'une fois à l'intérieur d'un même chemin. Les structures en boucle sont fréquemment utilisées dans les questionnaires pour modifier les réponses dont on a déterminé qu'elles étaient incorrectes. Par exemple, les questions d'ordre financier ou portant sur l'utilisation du temps peuvent être vérifiées au moyen de contrôles donnant lieu à un retour en arrière si la somme des questions composantes ne correspond pas au bon total.

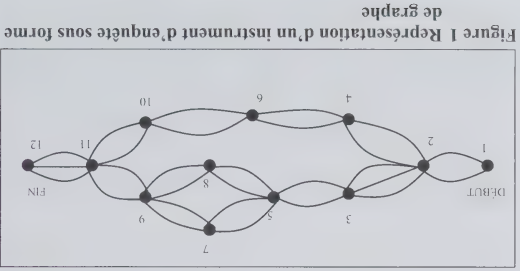


Figure 1 Représentation d'un instrument d'enquête sous forme de graphe

Application de la théorie des graphes à l'élaboration et à l'essai d'instruments d'enquête

Steven Elliott¹

Résumé

La présente étude porte sur l'application de la théorie des graphes à l'élaboration et à l'essai d'instruments d'enquête. Une approche fondée sur la théorie des graphes présente plusieurs avantages par rapport aux approches conventionnelles au chapitre de la structure et des caractéristiques du système de spécifications des instruments de recherche, en particulier les instruments assistés par ordinateur et de vaste portée. La possibilité de vérifier la connectivité de tous les éléments constitue l'un de ces avantages, de même que la capacité de simuler un instrument. Une telle approche permet aussi de produire des mesures servant à décrire l'instrument, par exemple le nombre de chemins et de parcours. Le concept de la « base » est abordé dans le contexte des essais des logiciels. Ce terme désigne le plus petit ensemble de parcours, parmi tous ceux que comporte un instrument, qui couvre tous les appartenements de liens et de nœuds. Ces parcours peuvent être utilisés à titre d'ensemble complet et économique de cas types pour faire l'essai des instruments.

Mots clés : Théorie des graphes ; interview assistée par ordinateur (IAO) ; élaboration de questionnaires ; essai de logiciel ; essai portant sur la base ; cas types.

1. Introduction

La théorie des graphes est une théorie mathématique relative aux ensembles de nœuds et de liens. Le terme « graphe » désigne la représentation visuelle d'un tel ensemble. Les graphes ont été utilisés dans de nombreux domaines d'étude pour modéliser des phénomènes de la réalité. Les premiers exemples ont trait à l'analyse de la logistique des transports (Berge 1976, page VII). Dans le cadre de telles analyses, une approche fondée sur la théorie des graphes sera utile pour déterminer entre autres les parcours offrant une efficacité maximale par rapport à différents endroits. Dans le graphe, ces endroits sont représentés par les nœuds, tandis que les liens représentent les chemins allant d'un endroit à l'autre.

Il existe aussi des applications de la théorie des graphes dans le domaine de la méthodologie d'enquête. Si l'on représente les questions d'un questionnaire d'enquête sous forme de nœuds et les chemins engendrés par les enchaînements de questions sous forme de liens, il devient possible de recourir à un graphe pour modéliser un questionnaire. Dès lors, bon nombre des théorèmes et des mesures descriptives associées à la théorie des graphes peuvent s'appliquer au regard des questionnaires. De plus, une approche fondée sur la théorie des graphes sera utile dans l'optique des processus de documentation et d'essai d'instruments d'enquête. Par exemple, un système de documentation contenant un tableau pour les questions et un autre pour les options de réponse permet de vérifier la connectivité de tous les éléments des instruments ainsi que de procéder à des simulations d'instruments fonctionnels. Une procédure d'essai où l'ensemble de cas types couvre à tout le moins la « base » d'un graphe d'instrument garantit que toutes les

combinaisons de liens et de nœuds font l'objet d'un essai à partir du plus petit nombre possible de cas.

Aux fins de l'élaboration, de la documentation et de l'essai de la plupart des instruments d'enquête, il n'est pas nécessaire de recourir à une représentation fondée sur la théorie des graphes. En effet, la plupart du temps, les instruments d'enquête comportent un nombre relativement peu élevé de questions et ne comptent pas beaucoup de points d'embranchement. Pensons notamment aux enquêtes sur la satisfaction de la clientèle et aux courtes enquêtes avec papier et crayon, comme le recensement américain. Pour ce genre d'instruments, les procédures conventionnelles de documentation et d'essai sont appropriées. Par contre, une approche reposant sur la théorie des graphes pourrait avoir sa place dans le cas des enquêtes complexes et de grande envergure, comme il y en a beaucoup à l'heure actuelle. Par exemple, en 2009, on a procédé à l'Enquête canadienne sur les capacités financières (ECCF) pour déterminer les connaissances et les comportements des Canadiens lorsqu'il est question de la prise de décisions financières ; les interviews téléphoniques assistées par ordinateur ont porté sur douze sections dont chacune comportait environ douze questions (Statistique Canada 2010). Menons encore les interviews trimestrielles (IPAO) de la Consumer Expenditure Survey (2010), enquête menée par le Bureau of Labor Statistics du département américain du Travail. Cette enquête compte 22 sections, dont la plupart comptent au moins trois sous-sections ; le nombre de questions par sous-section va de six à 90 (US Bureau of Labor Statistics 2010). Dans l'un et l'autre cas, une approche envisageable aux fins de documentation et d'essai.

1. Steven Elliott, Westat Incorporated, 1500 Research Blvd., Rockville, MD 20850-3158, États-Unis. Courriel : sdel Elliott2@verizon.net.

Nigimi, M. (1999). I've got your Number. *Journal of Accountancy*, 187(5), 79-83.

Porras, J., et English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. Dans *Proceedings of the Survey Research Method Section*. American Statistical Association, 4223-4228.

Saville, A. (2006). Using Benford's law to predict data error and fraud - An examination of companies listed on the JSE Securities Exchange. *South African Journal of Economic and Management Sciences*, 9(3), 341-354.

Schäfer, C., Schräpler, J., Müller, K. et Wagner, G. (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. *Schnollers Jahrbuch*, 125, 183-193.

Schnell, R. (1991). Der einfluss gefälschter Interviews auf survey ergebnisse. *Zeitschrift für Soziologie*, 20(1), 25-35.

Schräpler, J. et Wagner, G. (2003). Identification, Characteristics and Impact of Faked Interviews in Surveys - An analysis by means of genuine fakes in the raw data of SOEP. Document de discussion IZA séries, 969.

Schreiner, I., Pennie, K. et Newbrough, J. (1988). Interviewer falsification in census bureau surveys. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 491-496.

Scott, P., et Fasil, M. (2001). Benford's law: An empirical investigation and a novel explanation. Rapport technique de la CSM, Department of Computer Science, University Essex.

Stokes, L., et Jones, P. (1989). Evaluation of the interviewer quality control procedure for the post-enumeration survey. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 696-698.

Swanson, D., Cho, M. et Eitinge, J. (2003). Detecting possibly fraudulent data or error-prone survey data using Benford's law. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 4172-4177.

Thompson, S. (2010). Cluster Analysis for Anomaly Detection in Accounting Data. Collected Papers of the Nineteenth Annual Strategic and Emerging Technologies Research Workshop San Francisco, California.

Turner, C., Gribbe, J., Al-Tayyip, A. et Chromy, J. (2002). Falsification in Epidemiologic Surveys: Detection and Remediation (Ebauche de prépublication). Papier technique sur l'Health and Behavior Measurement. Washington DC : Research Triangle Institute. No. 53.

Annexe

Résultats des classifications automatiques fondées sur les équations 3 et 4 pour toutes les combinaisons possibles de classes

[illegible]

¹ Combinaison d'indicateurs non Pareto-donnée.
² Les valeurs moyennes de classe n'ont pas permis de déterminer la classe « à risque ».

Remerciements

Nous tenons à souligner l'appui financier de la Deutschen Forschungsgemeinschaft par la voie du projet « PP 1292: *Survey Methodology* ». Nous remercions en outre John Bushery et quatre examinateurs anonymes de leurs commentaires constructifs concernant notre article.

Bibliographie

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(1), 551-572.

Biemer, P., et Stokes, S. (1989). The optimal design quality control sample to detect interviewer cheating. *Journal of Official Statistics*, 5(1), 23-29.

Bushery, J., Reichert, J., Albright, K. et Rossiter, J. (1999). Using date and time stamps to detect interviewer falsification. Dans *Proceedings of the Survey Research Method Section, American Statistical Association*, 316-320.

Diekmann, A. (2002). Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung. Rapport technique manuscrit 06/2002, Institut für Technikfolgenabschätzung (ITfA), Wien.

Donoho, S. (2004). Early detection of insider trading in option markets. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 420-429.

Eyermann, J., Murphy, J., McCue, C., Hottlinger, C. et Kennel, J. (2005). Dépistage de la falsification des données par l'Internet : vécu par l'exploration de données. Dans le Recueil : Symposium (2005). *Les méthodologiques reliés aux besoins futurs d'information*. Statistique Canada.

Forman, G., et Schreiner, I. (1991). The design and analysis of
reinterview: An overview. Dans *Measurement Errors in Surveys*,
(Eds., P.B. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz
et S. Sudman), New York: John Wiley & Sons, Inc, 279-301.

Guterbock, T.M. (2008). *Falsification. Dans Encyclopedia of Survey Research Methods*, (Ed., P.J.Lavrakas), Sage Publications, Thousand Oaks, 1, 267-270.

Hardle, W., et Simar, L. (2007). *Applied Multivariate Statistical Analysis*, 2^e Edition. Springer, Berlin.

Hill, T. (1995). A statistical derivation of the significant digit law. *Statistical Science*, 10(4), 354-363.

Hill, T. (1999). The difficulty of faking data. *Chance*, 26, 8-13.

Hood, C., et Bushery, M. (1997). Getting more bang from the reinterviewer buck: Identifying 'At risk' interviewers. *Dans Proceedings of the Survey Research Method Section, American Statistical Association*, 820-824.

L.I., Brick, J., Tran, B. et Singer, P. (2009). Using statistical models for sample design of a reinterview program. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 4681-4695.

Murad, U., et Pinkas, G. (1999). Unsupervised Profiling for Identifying Superimposed Fraud. Lecture Notes in Computer Science, 1704, 251-261.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D. et Kennet, J. (2004). A system for detecting interviewer falsification. Article présenté à l'American Association for Public Opinion Research 59th Annual Conference.

Nigrini, M. (1996). A taxpayers compliance application of Benford's law. *Journal of the American Taxation Association*, 18, 72-91.

Tableau 6
Coefficients estimés standardisés et non standardisés (analyse

Variable	Coefficient (non standardisé)	Coefficient (standardisé)
Non-réponse partielle	0,767	0,917
Réponses « Autre »	-0,025	-0,129
Réponses extrêmes	0,075	0,821
Valeur de χ^2	-0,092	-0,562
Constante	-4,250	
Lambda de Wilks (Prob > F)		0,0254

4. Conclusion

Les données d'enquête peuvent être affectées par les intervieweurs qui fabriquent des données. La fabrication de données est un problème qu'il ne faut pas négliger, car il peut causer des biais importants. Même une petite quantité de données contrefaites peut altérer gravement les résultats des analyses empiriques ultérieures. Nous étendons les approches antérieures en vue de repérer les intervieweurs à risque en combinant plusieurs indicateurs dérivés directement des données d'enquête par classification automatique. Afin de démontrer notre approche, nous l'appliquons à un petit ensemble de données qui a été fabriqué en partie par des falsificateurs. Le fait que nous sachions dès le départ quels sont les falsificateurs nous permet d'évaluer les résultats de la classification automatique et de procéder par après à une analyse discriminante pour révéler dans quelle mesure les deux groupes d'intervieweurs peuvent être bien séparés au moyen des variables indicatrices. Des classifications automatiques de divers types sont effectuées. Toutes donnent lieu à la détermination d'une classe d'intervieweurs à risque, le ratio de non-réponse partielle et le ratio de réponses « Autre » étant les deux indicateurs les plus clairs. Nous arrivons pas à identifier parfaitement les falsificateurs. Cependant, dans tous les cas, la classe des intervieweurs à risque contient une part nettement plus élevée de falsificateurs que la deuxième classe. Les avantages de la classification automatique tiennent au fait que l'on obtient une classification nette des intervieweurs qui sont à risque et des autres intervieweurs, ce qui n'est pas le cas quand des indicateurs tels que la valeur de χ^2 sont examinés individuellement. En outre, elle nous permet de combiner l'information provenant de plusieurs indicateurs. En étudiant la performance de tous les sous-ensembles possibles d'indicateurs, nous constatons qu'en général, un grand nombre d'indicateurs permet de mieux repérer les falsificateurs. Le fait que diverses méthodes de classification produisent des résultats différents ne devrait pas être nécessairement considéré comme un défaut de notre approche. Selon la pondération choisie du coût d'un falsificateur non détecté comparativement à celui d'une fausse alarme, on pourrait en dernière

L'application à un petit ensemble de données démontre un autre mérite de notre approche : elle a été testée et donne de bons résultats dans une situation dans laquelle le nombre de questionnaires par intervieweur était assez limité (trois des falsificateurs n'ont soumis que dix questionnaires). Si un petit nombre de questionnaires par intervieweur est suffisant pour effectuer l'analyse, on pourrait aussi imaginer de la mettre en œuvre durant la période principale de travail sur le terrain, quand les intervieweurs n'ont soumis qu'un certain pourcentage de leurs questionnaires. Les falsificateurs pourraient alors être remplacés par d'autres intervieweurs qui mèneront l'enquête auprès des unités qui auraient dû être interviewées par les falsificateurs. Évidemment, en examinant nos résultats, nous ne devons pas perdre de vue que nous avons appliqué notre méthode à un ensemble de données dans lequel une forme très grave de falsification des données a eu lieu : d'une part, nous avons des falsificateurs qui ont contrefait les données de tous leurs voisins (responsables) complètement et, d'autre part, nous avons des intervieweurs qui (on le présume) ont effectué leur travail honnêtement, ce qui facilite la discrimination entre les intervieweurs honnêtes et les intervieweurs malhonnêtes. En outre, la taille de notre échantillon, qui ne comprend que 13 intervieweurs, est assez limitée. Il serait intéressant d'explorer l'utilité de notre approche lorsqu'on l'applique à de plus grands ensembles de données, étant donné que la part d'interviews falsifiées dans les grandes enquêtes s'avère plus faible que dans notre cas. Qui plus est, les grands ensembles de données pourraient permettre de construire des indicateurs supplémentaires pour la classification automatique. Si l'enquête comprend un programme de réinterview, il serait possible d'évaluer l'utilité de notre approche en comparant le « succès » d'une réinterview aléatoire avec celui d'une réinterview axée sur les intervieweurs considérés comme étant à risque. Nous avons également l'intention de poursuivre l'analyse dans des conditions expérimentales. Des conditions appropriées permettent de s'assurer que l'on obtient un ensemble de données qui a été recueilli partiellement en menant des interviews réelles et partiellement fabriquées en disant à certains participants à l'expérience de remplir le questionnaire eux-mêmes.

dans quelle mesure les indicateurs employés dans l'analyse permettent de bien séparer les divers groupes et si l'appartenance à un groupe peut être prédite correctement [voir Hårdle et Sîmar (2007) pour une introduction à l'analyse discriminante]. Dans une analyse discriminante linéaire, les coefficients b_0 et b_j de la fonction discriminante $D = b_0 + \sum_{j=1}^p b_j x_j$ sont déterminés de façon telle qu'ils maximisent l'écart entre les valeurs D moyennes des deux groupes distincts et simultanément diminuent avec les écarts entre les valeurs D des éléments à l'intérieur des groupes. Dans notre cas, les x_j sont nos quatre variables indicatrices et nous obtenons deux groupes en séparant les faisiolateurs et les intervieweurs honnêtes. Nous utilisons les probabilités a priori correspondant à la taille de groupe relative (4/13 et 9/13) afin de prédire l'appartenance à un groupe. Le tableau 5 donne les résultats. Manifestement, les quatre variables permettent d'obtenir une bonne séparation des faisiolateurs et des intervieweurs honnêtes, car l'appartenance à un groupe est prédite correctement dans tous les cas sauf un.

0,71. Le lambda de Wilks de l'analyse discriminante est statistiquement significatif au seuil de signification de 5 %.

Tableau 5

Intervieweur	Groupe prédit	Groupe réel	Fonction discriminante
F1	1	1	-2,878
F2	1	1	-3,376
F3	2	1	-0,541
F4	1	1	-1,955
H1	2	2	1,828
H2	2	2	1,060
H3	2	2	1,747
H4	2	2	1,616
H5	2	2	0,706
H6	2	2	0,777
H7	2	2	-0,041
H8	2	2	1,765
H9	2	2	-0,710

possibles d'indicateurs, y compris les cas ne s'appuyant que sur un seul indicateur. Les résultats (voir le tableau 7 en annexe) indiquent généralement que l'augmentation du nombre d'indicateurs améliore les résultats. Toutefois, il existe aussi des combinaisons comportant un plus petit nombre d'indicateurs qui produisent des résultats semblables à ceux fondés sur les quatre indicateurs pris ensemble. Pour déterminer quelle combinaison d'indicateurs est la meilleure, il faudrait s'appuyer sur l'établissement hautement subjectif du coût relatif de la non-détection d'un faussificateur comparativement à celui d'une fausse alarme. Toutefois, on peut déterminer quelles combinaisons d'indicateurs sont non Pareto-dominées en ce sens qu'il n'existe aucune autre combinaison présentant moins de faussificateurs non décelés (fausses alarmes) et en même temps ne présentant pas plus de fausses alarmes (faussificateurs non décelés). La combinaison d'indicateurs constituée des quatre indicateurs est la seule qui est non Pareto-dominée, quelle que soit l'équation utilisée. Par contre, les combinaisons ne comprenant qu'un seul indicateur sont Pareto-dominées dans six cas sur huit.

Tableau 3
Résultats des trois méthodes de classification employées

Classification hiérarchique								
Intervaleur	F1	F2	F3	F4	H1	H2	H3	H4
Classe	1	1	2	2	2	2	2	2
Distance entre les classes divisée par la distance intra-classe								
Intervaleur	F1	F2	F3	F4	H1	H2	H3	H4
Classe	1	1	2	1	2	2	2	2
Distance entre les éléments d'une classe								
Intervaleur	F1	F2	F3	F4	H1	H2	H3	H4
Classe	1	1	1	1	1	1	1	1
Intervaleur	F1	F2	F3	F4	H1	H2	H3	H4
Classe	1	1	1	1	1	1	1	1

Tableau 4
Moyennes des variables indicatrices par classe pour les trois compositions des classes

Valeur de χ^2	Réponses	« Autre »	Réponses	Non-réponse	partielle
12,43	1	2	1	2	1
25,55	2	1	2	1	2
42,60	1	2	1	2	1
61,37	2	1	2	1	2
7,64	1	2	1	2	1
0,22	2	1	2	1	2
2,32	1	2	1	2	1
0,86	2	1	2	1	2
Moyenne	12,43	25,55	42,60	61,37	7,64
Distance entre les classes	1	2	1	2	1
Distance entre les classes divisée par la distance intra-classe	1	2	1	2	1
Moyenne	12,43	25,55	42,60	61,37	7,64
Distance entre les éléments d'une classe	1	2	1	2	1
Moyenne	11,73	21,43	50,83	60,92	9,06
Classe	1	2	1	2	1
Moyenne	0,68	2,80	0,92	1	2

3.3 Analyse discriminante

Enfin, nous nous tournons vers l'analyse discriminante pour vérifier si les hypothèses concernant le comportement des falsificateurs sont valides. L'analyse discriminante peut être appliquée si les classes sont connues afin d'évaluer

entre toutes les paires possibles d'éléments dans la même classe. En divisant cette somme par le nombre de paires possibles, on obtient la distance intra-classe moyenne.

Le tableau 3 donne les résultats des trois méthodes de classification. Dans l'analyse hiérarchique avec lien entre groupes, les trois falsificateurs F1, F2 et F4 forment la classe 1, et le falsificateur F3 ainsi que tous les intervieweurs honnêtes, la classe 2. Donc, nous arrivons à séparer les deux groupes d'intervieweurs, à l'exception d'un falsificateur. Cependant, si nous ne savions pas au départ quels intervieweurs ont fabriqué des données et lesquels sont honnêtes, nous devrions décider laquelle des deux classes contient les intervieweurs à risque. Cela peut se faire en comparant les moyennes des variables indicatrices dans chaque classe présentées au tableau 4. Pour la classification hiérarchique, les moyennes pour le ratio de non-réponses partielles et pour le ratio de réponses « Autre » sont manifestement plus faibles dans la classe 1. Il en est de même de la moyenne pour le ratio de réponses extrêmes, quoique la différence entre les deux classes soit moins frappante. Enfin, une moyenne plus élevée de la valeur de χ^2 peut être observée pour la classe 1. Étant donné ces résultats, en nous en tenant aux hypothèses susmentionnées concernant le comportement des falsificateurs, nous déterminons cor-

rectement que la classe 1 est celle qui contient tous les intervieweurs à risque. Nous avons également essayé d'améliorer les résultats de la classification hiérarchique en utilisant les moyennes de classe présentées au tableau 4 comme point de départ pour la classification par la méthode des K-moyennes. Cependant, l'application de l'algorithme des K-moyennes n'a donné lieu à aucune modification de la composition des classes.

La composition des classes qui maximise l'équation (3) est identique à celle obtenue en utilisant la classification hiérarchique. Par conséquent, comme le montre le tableau 4, les moyennes des indicateurs dans les deux classes sont également identiques.

La composition des classes qui minimise l'équation (4) est légèrement différente. La classe 1 contient maintenant tous les falsificateurs et un intervieweur honnête. Les moyennes des variables indicatrices indiquent de nouveau clairement que la classe 1 est celle qui contient les intervieweurs à risque. Ce résultat est très satisfaisant. Tous les falsificateurs sont répétés et une seule fausse mesure de distance, qui cette fois doit être minimisée, est calculée comme il suit :

$$(3) \quad \frac{\sum_{i=1}^f (\bar{d}_{1i}^2 - \bar{d}_{2i}^2)}{\sum_{i=1}^f \bar{d}_{1i}^2 + \sum_{i=1}^f \bar{d}_{2i}^2}.$$

Pour examiner toutes les compositions de classes possibles qui requiert d'importants calculs quand le nombre d'intervieweurs augmente. À la place, on pourrait recourir à l'approche consistant à essayer toutes les compositions de classes et de choisir la meilleure en fonction d'une certaine fonction cible. (L'analyse a été exécutée en MATLAB et le code du programme peut être obtenu sur demande.) Cette procédure est clairement supérieure à la classification hiérarchique, car elle fait en sorte que soit déterminée la composition globalement optimale des classes. Cependant, nous donnons aussi les résultats de la classification hiérarchique, car elle est assez pratique comparativement à la composition globale optimale.

Il serait également intéressant de voir quelle composition optimale des classes on obtiendrait si, au lieu de maximiser l'équation (3), on minimisait la moyenne du carré de la distance euclidienne entre toutes les paires possibles dans une classe. En fait, cette idée ressemble fort à la fonction cible pertinente dans les procédures de classification hiérarchique présentées plus haut. Notre deuxième mesure de distance, qui cette fois doit être minimisée, est calculée

$$(4) \quad \frac{\sum_{i=1}^f \sum_{j=1}^f CDE_{ij}^k + \sum_{i=1}^f \sum_{j=1}^f CDE_{ij}^k}{\sum_{i=1}^f \sum_{j=1}^f CDE_{ij}^k + \sum_{i=1}^f \sum_{j=1}^f CDE_{ij}^k}.$$

Le numérateur est égal à la somme des distances euclidiennes entre les éléments i et j , calculée comme $CDE_{ij}^k = \sum_{i=1}^f (d_{ij}^k)^2$.

concernant les caractéristiques du ménage, la richesse en ressources, ainsi que le revenu et les dépenses. La plupart des questions étaient de type fermé. Quelques-unes seules métriques ont été recueillies pour les dépenses des ménages, comme le loyer ou l'achat de terrains, de semences, ou d'engrais, ou les impôts, ainsi que pour le revenu du ménage provenant de diverses sources, comme le travail autonome agricole et non agricole, et les transferts publics et privés.

Au moment où les interviews de l'enquête de 2007 ont été menées, aucun des chercheurs allemands n'était présent dans les villages. Les questionnaires ont été recueillis immédiatement après la réalisation de l'enquête dans le premier village. Un premier examen de ces questionnaires a suscité des soupçons, parce que le papier des questionnaires paraissait très propre et très blanc. Le papier ne présentait pas de salissures ni de coins cornés. En comparant les réponses figurant sur divers questionnaires soumis par un intervieweur, nous avons découvert deux questionnaires contenant des réponses identiques. Étant donné que nous avions demandé d'indiquer le montant du revenu en provenance de diverses sources en valeur métrique, il était fort peu probable que les réponses sur les deux questionnaires soient identiques. Comme nous n'obtenions aucune explication de la part des partenaires du projet, nous avons réinterviewé sur place un sous-échantillon de 10 % de l'échantillon original. Aucun des ménages réinterviewés n'a déclaré qu'il avait déjà été interviewé. Après que nous ayons détecté la fabrication des interviews, les partenaires ont reconnu que toutes les interviews avaient été contrefaites. Nous avons naturellement cessé de travailler avec tous les intervieweurs et partenaires, et créé un nouveau groupe de recherche local.

En février 2008, l'enquête a été répétée dans le même pays. Comme nous l'avons mentionné plus haut, nous avons sélectionné de nouveaux villages et ménages conformément aux critères susmentionnés. Nous avons recruté neuf étudiants pour les interviews et organisé une supervision sur place durant l'enquête. Dans la plupart des cas, les interviews ont eu lieu dans une école ou à l'hôtel de ville afin que nous puissions surveiller tous les intervieweurs. Quand les interviews ont eu lieu au domicile des familles participant à l'enquête, nous avons assisté à certaines d'entre elles. Étant donné cette procédure, nous présumons que les réponses au questionnaire de l'enquête de 2008 n'ont pas été contrefaites.

Le présent article porte sur un total de 250 interviews de ménages réalisées par 13 intervieweurs, dont quatre étaient des falsificateurs de l'enquête de 2007 (les interviews soumises par l'un d'eux ont été exclues, car il n'a remis que trois questionnaires) qui ont définitivement contrefait les

Les données sur lesquelles porte la présente étude proviennent d'enquêtes-ménages menées en novembre 2007 et en février 2008 dans un pays de la Communauté des États indépendants (CEE) (ancienne Union soviétique). L'enquête a été réalisée dans le cadre d'un projet de recherche international sur les réformes agraires et la pauvreté rurale. Nous avons l'intention d'interviewer 200 ménages dans quatre villages en 2007. Après avoir déterminé que toutes les interviews avaient été contrefaites dans le premier village étudié, nous avons interrompu l'enquête et lancé un nouveau cycle avec de nouveaux intervieweurs dans d'autres villages en février 2008. Tous les villages ont été sélectionnés en s'appuyant sur des critères qualitatifs tels que la structure de la production agricole et la mise en œuvre de réformes agraires. Dans chaque village, les ménages ont été échantillonnés aléatoirement en se servant de listes de ménages fournies par les maires des villages. Non seulement cette procédure assurait que tous les ménages soient sélectionnés au hasard, mais elle fournissait aussi le fondement pour des réinterviews, puisque tous les ménages étaient définis exactement. Cependant, ces nouvelles interviews n'avaient pas été planifiées au tout début de l'enquête. Comme les ménages possédaient rarement un téléphone, les appels de vérification n'étaient pas possibles et les nouvelles interviews de ces ménages ont nécessité des déplacements dans le village pour procéder à la réinterview sur place, ce qui a entraîné d'importantes dépenses d'argent et de temps. Cinq intervieweurs ont été recrutés au moment de la première enquête en 2007. Deux d'entre eux étaient les partenaires locaux du projet de recherche. Ils avaient participé à l'élaboration du questionnaire et étaient responsables de la coordination des enquêtes dans leur pays. Les trois autres intervieweurs étaient des étudiants engagés par les partenaires. Le questionnaire comprenait diverses sections

3.1 Sources des données

3. Résultats

toujours disponible.

interviews dans lesquelles sont surpondérés les cas ayant une probabilité élevée d'être frauduleux selon le modèle de régression logistique donnent lieu à la détection d'un plus grand nombre de cas réels de fabrication de données que les échantillons sélectionnés purement au hasard. Cependant, il est évident qu'il faut disposer de données de réinterviews antérieures pour lesquelles l'état de falsification est connu pour exécuter la régression logistique.

Le nombre de combinaisons de réponses rares ou improbables dans les questionnaires soumis par un intervieweur (Murphy et coll. 2004 ; Porras et English 2004) et la comparaison de la composition ou des statistiques descriptives des ménages dans les questionnaires d'un intervieweur avec les données pour l'échantillon complet (Turner et coll. 2002 ; Murphy et coll. 2004) sont d'autres indicateurs discutés dans la littérature.

Un autre moyen de déceler les données contrefaites devenu populaire ces dernières années est l'utilisation de la loi de Benford (Schäppler et Wagner 2003 ; Swanson et coll. 2003 ; Porras et English 2004 ; Schäfer et coll. 2005), dont nous discuterons à la section 2, y compris le succès avec lequel elle a permis de détecter les interviewers contrefaits dans le cadre d'études antérieures. La section 2 décrit aussi notre approche statistique en vue de dépister les falsificateurs. La section 3 présente les données sur lesquelles s'appuie notre analyse ainsi que nos résultats. Nous concluons l'article par une discussion de nos résultats.

Quand le physicien Frank Benford a remarqué que les pages des tables de logarithmes contenant les logarithmes des nombres faibles (1 et 2) étaient plus souvent utilisées que celles contenant les logarithmes de nombres plus élevés (8 et 9), il a commencé à étudier la distribution du premier chiffre d'une grande gamme de types de nombres, comme ceux figurant à la première page d'un journal, dans les adresses de voiture ou dans les poids moléculaires (Benford 1938). Benford a constaté que la distribution du premier chiffre non nul des nombres pouvait être décrite par la formule qui suit que l'on a appelée « loi de Benford » :

$$\text{Prob}(\text{premier chiffre} = d) = \log_{10} \left(1 + \frac{d}{1} \right) \quad (1)$$

Toutefois, les séries de nombres étudiées par Benford (1938) ne semblaient pas se conformer toutes à cette loi. Par conséquent, la question qui se posait était de savoir quelle sorte de données on pouvait supposer produire des premiers chiffres de nombre obéissant à la loi. Des discussions de cette question peuvent être consultées dans Hill (1995), Nigriti (1996), Hill (1999), et Scott et Fashi (2001). La

2. Méthodes

2.1 Loi de Benford

détection de la fraude financière est un domaine dans lequel l'application de la loi de Benford s'est beaucoup répandue durant la dernière décennie (Nigriti 1996 ; 1999 ; Saville 2006). Bien que les résultats de ces études ne soient pas pertinents dans notre contexte, il est intéressant de mentionner que le consensus dans la littérature semble être que l'on peut supposer que les valeurs monétaires suivent la loi de Benford. Swanson et coll. (2003) montrent que la distribution des premiers chiffres des nombres dans la Consumer Expenditure Survey des États-Unis est proche de la distribution de Benford.

La notion fondamentale qui sous-tend l'utilisation de la loi de Benford pour détecter les données contrefaites est que les falsificateurs ne connaissent vraisemblablement pas la loi ou qu'ils ne sont pas capables de fabriquer des données qui la suivent. Par conséquent, un écart important de la distribution des premiers chiffres par rapport à la distribution de Benford dans un ensemble de données indique que les données pourraient être contrefaites. Naturellement, il faut se demander si la nature des données permet de supposer qu'elles suivent la loi de Benford si elles sont authentiques. La loi de Benford ne peut pas être appliquée si les questionnaires ne contiennent que très peu de variables métriques, voire aucune.

Schäppler et Wagner (2003), ainsi que Schäfer et coll. (2005) utilisent la loi de Benford pour déceler la fabrication des données dans le GSOEP. Dans les deux études, tous les questionnaires administrés par chaque intervieweur sont combinés et vérifiés afin de déterminer si la distribution des premiers chiffres des nombres figurant dans les questionnaires s'écarte de manière significative de la loi de Benford. Cela peut se faire en calculant la statistique χ^2 :

$$\chi^2_i = n_i \sum_{j=1}^p \frac{h_{jd}^2}{(h_{jd}^2 - h_{jd}^2)} \quad (2)$$

Dans les deux études, les valeurs critiques de χ^2 sont supposées dépendre de la taille de l'échantillon n et sont par conséquent corrigées pour ce paramètre. Les résultats obtenus semblent prometteurs. L'ajustement de la distribution des premiers chiffres des nombres à la loi de Benford est en général nettement moins bon pour les questionnaires des

souligner que cette connaissance a priori n'est pas une condition préalable à l'emploi de la méthode.

La question de l'identification des intervieweurs à risque a été abordée durant les années 1980, mais la littérature Census Bureau a mis en œuvre l'interviewer falsification Study. En s'appuyant sur l'information recueillie dans le contexte de cette étude, Schreiner et coll. (1988) constatent que les intervieweurs ayant peu d'ancienneté sont plus susceptibles que les autres de fabriquer des données. Hood et Bushery (1997) utilisent plusieurs indicateurs pour de Health Interview Survey (NHIS). Par exemple, ils calculent le taux par intervieweur de ménages désignés comme étant inadmissibles ou de ménages n'ayant pas de numéro de téléphone, et comparent ces taux aux données de recensement pour les régions pertinentes. En cas d'écart important, l'intervieweur est signalé et une nouvelle interview est effectuée. Les taux de détection parmi les intervieweurs signalés s'avèrent plus élevés que ceux observés pour des échantillons aléatoires de cas à réinterviewer. Turner et coll. (2002) constatent aussi, en examinant les données de la Baltimore STD and Behaviour Survey, que les intervieweurs qui fabriquent des données indiquent moins fréquemment des numéros de téléphone que les intervieweurs honnêtes. Dans le cas de l'interview assisté par ordinateur, Bushery, Reichert, Albright et Rossiter (1999), ainsi que Murphy et coll. (2004) proposent d'utiliser l'horodatage, c'est-à-dire l'enregistrement de l'heure et de la durée de l'interview par l'ordinateur, pour découvrir les intervieweurs suspects. Ceux qui ont besoin d'un temps anormalement long ou court pour administrer le questionnaire complet ou certains modules, ou ceux qui administrent un nombre étonnamment grand de questionnaires durant une période donnée pourraient alors être signalés comme des intervieweurs à risque. Schaffer et coll. (2005) supposent que les falsificateurs évitent les réponses extrêmes lorsqu'ils fabriquent les données. D'après des données du GSOEP, ces auteurs calculent la variance des réponses pour chaque question de tous les questionnaires administrés par un intervieweur et totalisent ces variances. Grâce à d'autres mécanismes de contrôle intégrés dans le GSOEP, les falsificateurs sont connus et il s'avère qu'on les retrouve parmi les intervieweurs ayant les variances globales les plus faibles. Porras et English (2004) adoptent une approche similaire et constatent aussi que les falsificateurs produisent des variances plus faibles que celles observées pour les questionnaires remplis honnêtement. Li et coll. (2009) combinent plusieurs indicateurs prédictifs dans un modèle de régression logistique dans lequel l'état de falsification connu d'un intervieweur sert de variable dépendante binaire. Les auteurs constatent que les échantillons de nouvelles

dans les questionnaires pour définir un groupe d'intervieweurs à risque. L'idée n'est pas nouvelle ; on trouve dans la littérature spécialisée plusieurs exemples de ce genre d'approche (Hood et Bushery 1997 ; Diekmann 2002 ; Turner, Gribbe, Al-Tayyib et Chorny 2002 ; Schröpler et Wagner 2003 ; Swanson, Cho et Eltinge 2003 ; Murphy, Baxter, Eyerman, Cunningham et Kennet 2004 ; Porras et English 2004 ; Schaffer, Schröpler, Müller et Wagner 2005 ; Li, Brick, Tran et Singer 2009). Cependant, à l'exception des travaux de Li et coll. (2009), les tests effectués dans ces études reposent sur l'examen d'indicateurs unilatéraux des données des intervieweurs pour déceler les falsificateurs. Certaines études comprennent le calcul de plusieurs indicateurs, mais en les considérant tous séparément. Nous combinons plusieurs indicateurs dans des analyses par classification automatique, ce qui permet d'obtenir une meilleure classification des falsificateurs éventuels comparativement aux approches antérieures. Autant que nous sachions, cette procédure est une innovation dans le contexte du dépistage des intervieweurs qui falsifient les données, mais elle a déjà été employée dans d'autres domaines afin de déceler des comportements frauduleux. L'idée fondamentale est que les caractéristiques des « cas » frauduleux (ce qui définit un cas dépend du contexte) présentent, comparativement aux cas honnêtes, des schémas frappants qu'il est possible de révéler si ces caractéristiques sont considérées simultanément dans une classification automatique. Murad et Pinkas (1999) essayent de déceler la fraude dans l'industrie des télécommunications par classification des profils d'appels des clients. Un appel est caractérisé par plusieurs indicateurs, dont l'heure de l'appel ou la destination de celui-ci. Thiprungrat (2010) regroupe les demandes d'indemnisation en exécution d'un contrat d'assurance-vie collective soumises par les clients des compagnies d'assurance-vie en s'appuyant sur plusieurs caractéristiques des demandes. Celles qui forment de très petits groupes sont considérées comme étant suspectes. Donoho (2004) utilise la classification automatique, entre autres, pour dégager les tendances des marchés des options susceptibles de révéler des opérations d'initié.

Nous disposons d'un petit ensemble de données d'enquête (voir la sous-section 3.1 pour une description plus détaillée de cet ensemble) qui est constitué en partie de données falsifiées. Issu de 250 questionnaires administrés par 13 intervieweurs, l'ensemble de données est de taille assez limitée et la mesure dans laquelle nos résultats peuvent être généralisés à de plus grands ensembles de données n'est pas claire. Toutefois, cet ensemble de données nous permet d'illustrer notre approche. Le fait que nous sachions quelles données ont été recueillies honnêtement et lesquelles ont été contrefaites permet de procéder à une première évaluation de l'approche. Il convient toutefois de

Une approche statistique pour déceler la falsification des données d'enquête par les intervieweurs

Sebastian Bredl, Peter Winker et Kerstin Kötschau¹

Résumé

Les données d'enquête peuvent être falsifiées par les intervieweurs, la fabrication de données étant la forme de falsification la plus flagrante. Même un petit nombre d'interviews contrefaites peuvent fausser gravement les résultats d'analyses empiriques subséquentes. Outre l'exécution de réinterviews, certaines approches statistiques ont été proposées pour repérer ce genre de comportement frauduleux. À l'aide d'un petit ensemble de données, le présent article illustre comment la classification automatique, qui n'est ordinairement pas employée dans ce contexte, pourrait être utilisée pour repérer les intervieweurs qui falsifient les tâches qui leur sont assignées. Plusieurs indicateurs sont combinés pour classer les intervieweurs « à risque » en se fondant uniquement sur les données recueillies. Cette classification multivariée semble supérieure à l'utilisation d'un seul indicateur tel que la loi de Benford.

Mots clés : Fabrication de données ; falsificateur ; loi de Benford ; classification automatique.

1. Introduction

Lorsque la collecte des données se fait par interview, il convient de se préoccuper de la qualité des données. Cette dernière peut souffrir si le répondant fournit des réponses fausses ou imprécises ou que le questionnaire est mal conçu, ou encore si l'intervieweur s'écarte de la procédure d'interview établie. S'il le fait délibérément, on parle de « falsification des données par l'intervieweur » (Schreiner, Pennie et Newbrough 1988) ou de « tromperie » (Schräpler et Wagner 2003).

L'intervieweur peut falsifier les données de nombreuses façons (voir Gutierbock 2008). Des formes assez subtiles de falsification consistent à interviewer le mauvais membre du ménage ou à mener l'enquête par téléphone quand l'interview devrait avoir lieu sur place. La forme de falsification la plus grave est la fabrication d'interviews complètes sans jamais prendre contact avec le ménage concerné. Dans notre analyse, nous traitons de ce dernier cas.

Les interviews contrefaites peuvent avoir de graves répercussions sur les statistiques produites d'après les données d'enquête. Schnell (1991), ainsi que Schräpler et Wagner (2003) donnent des preuves que l'effet sur les statistiques univariées pourrait être moins prononcé, à condition que la proportion de falsificateurs demeure suffisamment faible et que la « qualité » des données contrefaites soit élevée. En revanche, même une faible proportion d'interviews contrefaites peut suffire à introduire un biais important dans les statistiques multivariées. Schräpler et Wagner (2003) constatent que l'inclusion de données contrefaites provenant du

panel socioéconomique de l'Allemagne (GSOEP) dans une régression multivariée réduit d'environ 80 % l'effet de la formation sur le logarithme de la rémunération brute, bien que la part d'interviews contrefaites soit inférieure à 2,5 %. Cet exemple montre qu'il est important de repérer ces interviews.

Le moyen le plus fréquent de déceler les intervieweurs qui falsifient les données consiste à procéder à une nouvelle interview (Biemer et Stokes 1989). Dans ce cas, un superviseur prend contact avec certains ménages qui auraient dû être interviewés afin de vérifier si l'intervieweur leur a effectivement rendu visite. Cependant, pour des raisons budgétaires, il est impossible de réinterviewer tous les ménages qui participent à une enquête (voir Forsman et Schreiner 1991). Par conséquent, il faut déterminer comment optimiser l'échantillon de nouvelles interviews de manière à déceler le mieux possible les falsificateurs. En général, il semble utile de sélectionner pour une nouvelle interview les ménages interrogés par un intervieweur qui, selon des caractéristiques associées aux réponses obtenues dans ces interviews, est plus susceptible que les autres de fabriquer des données. Dans ce contexte, Hood et Busberry (1997) utilisent le terme d'intervieweur « à risque ». Si la sélection des cas à réinterviewer se fait par échantillonnage en deux étapes, où les intervieweurs sont sélectionnés à la première étape et les enquêtes interrogés par ces intervieweurs, à la deuxième étape [comme le recommandent Forsman et Schreiner (1991)], les intervieweurs à risque pourraient être suréchantillonnés à la première étape.

Dans le présent article, nous montrons une approche purement statistique qui s'appuie sur les données contenues

1. Sebastian Bredl, Département de statistique et d'économétrie, Université Justus-Liebig, 35394 Gießen, Licher Straße 64, Allemagne. Courriel : sebastian.bredl@wirtschaft.uni-giessen.de ; Peter Winker, Département de statistique et d'économétrie, Université Justus-Liebig, 35394 Gießen, Licher Straße 64, Allemagne. Courriel : peter.winker@wirtschaft.uni-giessen.de ; Kerstin Kötschau, Hanse Parlament, 22587 Hamburg, Blankenburger Landstrasse 7, Allemagne. Courriel : kkoetschau@hanse-parlament.eu.



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 38, numéro 1, juin 2012

Table des matières

Articles réguliers

Sebastian Bredl, Peter Winker et Kerstin Kötschau	1
Une approche statistique pour déceler la falsification des données d'enquête par les intervieweurs.....	1
Steven Elliott	13
Application de la théorie des graphes à l'élaboration et à l'essai d'instruments d'enquête.....	13
G. Hussain Choudhry, J.N.K. Rao et Michael A. Hidiroglou	25
À propos de la répartition de l'échantillon pour une estimation sur domaine efficace.....	25
Ted Chang	33
Options de calage pour remplacer la poststratification de données croisées à deux dimensions.....	33
Paul Knottnerus et Arnout van Delden	45
À propos de la variance des variations estimées d'après des panels rotatifs et des strates dynamiques.....	45
Dan Liao et Richard Valliant	57
Facteurs d'inflation de la variance dans l'analyse des données d'enquêtes complexes.....	57
Hung-Mio Lin, Hae-Young Kim, John M. Williamson et Virginia M. Lesser	69
Estimation de coefficients d'accord d'après des données d'enquête.....	69
Jörg Drechsler et Jerome P. Reiter	81
Conjuguer des données synthétiques et le sous-échantillonnage pour créer des fichiers de microdonnées à grande diffusion pour les enquêtes à grande échelle.....	81
Balgorin Nandram et Myron Katzoff	89
Un modèle hiérarchique bayésien de non-réponse pour les données catégoriques d'un tableau à double entrée provenant de petits domaines avec incertitude au sujet de l'ignorabilité.....	89
Communications brèves	
Phillip S. Kott	103
Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes.....	103
Autres revues	109

Techniques d'enquête est répertoriée dans *The ISI Web of knowledge (Web of science)*, *The Survey Statistician*, *Statistical Theory and Methods Abstracts* et *Social Research Methodology*, *Erasmus University*. On peut en trouver les références dans *Current Index to Statistics*, et *Journal Contents in Qualitative Methods*. La revue est également citée par *SCOPUS* sur les bases de données *Elsevier Bibliographic Databases*.

COMITÉ DE DIRECTION

Président

J. Kovar

Anciens présidents

D. Royce (2006-2009)

G.J. Brackstone (1986-2005)

R. Platek (1975-1986)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.A. Hidiroglou, *Statistique Canada*

chef délégué

H. Mantel, *Statistique Canada*

Rédacteurs associés

J.-F. Beaumont, *Statistique Canada*

J. van den Brakel, *Statistics Netherlands*

J.M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

R. Chambers, *Centre for Statistical and Survey Methodology*

J.L. Eltinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistique Canada*

D. Haziza, *Université de Montréal*

B. Hultigier, *University of Applied Sciences Northwestern Switzerland*

D. Judkins, *Westat Inc.*

D. Kasprzyk, *NORC at the University of Chicago*

P. Kott, *National Agricultural Statistics Service*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallee, *Statistique Canada*

P. Lynn, *University of Essex*

D.J. Males, *National Center for Health Statistics*

G. Nathan, *Hebrew University*

J. Opsomer, *Colorado State University*

Rédacteurs adjoints

C. Bocci, K. Bosu, P. Dick, G. Dubreuil, S. Godbout, C. Leon, S. Matthews, Z. Patak, S. Rubin-Bleuer et Y. You, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et l'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'état de l'art et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféablement en Word au rédacteur en chef, rfc@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca).

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ C/A par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 \$ C/A (6 \$ x 2 exemplaires), autres pays, 20 \$ C/A (10 \$ x 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association internationale de statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.

Techniques d'enquête

Une revue
éditée

par Statistique Canada

Juin 2012 • Volume 38 • Numéro 1

Publication autorisée par le ministre responsable de Statistique Canada

© Ministère de l'Industrie, 2012

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division de la gestion de l'information, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 2012

N° 12-001-XPB au catalogue

Périodicité : semestrielle

ISSN 0714-0045

Ottawa

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Télécopieur	1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit n° 12-000-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Ce produit est aussi disponible en version imprimée standard au prix de 30 \$CAN l'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste
- En personne auprès des agents et librairies autorisés.
Ottawa (Ontario) K1A 0T6
150, promenade Tunney's Pasture
Immeuble R.-H.-Coats, 6^e étage

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».



Statistique
Canada

Statistics
Canada

Canada

Numéro 1

•

Volume 38

•

Juin 2012

Une revue
éditée
par Statistique Canada

N° 12-001-XPB au catalogue

Techniques d'enquête

Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

December 2012

•

Volume 38

•

Number 2



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access and order this product

This product, Catalogue no. 12-000-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and browse by "Key resource" > "Publications."

This product is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail
Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Survey Methodology

A journal
published by
Statistics Canada



December 2012 • Volume 38 • Number 2

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2012

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Information Management Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

December 2012

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

MANAGEMENT BOARD

Chairman J. Kovar
Past Chairmen D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Members G. Beaudoin
S. Fortier (Production Manager)
J. Gambino
M.A. Hidirolou
H. Mantel

EDITORIAL BOARD

Editor M.A. Hidirolou, *Statistics Canada*
Deputy Editor H. Mantel, *Statistics Canada*

Past Editor J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Associate Editors

J.-F. Beaumont, *Statistics Canada*
J. van den Brakel, *Statistics Netherlands*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
R. Chambers, *Centre for Statistical and Survey Methodology*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
D. Haziza, *Université de Montréal*
B. Hülliger, *University of Applied Sciences Northwestern Switzerland*
D. Judkins, *Westat Inc.*
D. Kasprzyk, *National Opinion Research Center*
J.K. Kim, *Iowa State University*
P.S. Kott, *RTI International*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
P. Lynn, *University of Essex*
D.J. Malec, *National Center for Health Statistics*
G. Nathan, *Hebrew University*

J. Opsomer, *Colorado State University*
D. Pfeiffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *Office for National Statistics*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *National Opinion Research Center*
C. Wu, *University of Waterloo*
W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard University*

Assistant Editors C. Bocci, K. Bosa, G. Dubreuil, C. Leon, S. Matthews, Z. Patak, S. Rubin-Bleuer and
Y. You, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

Survey Methodology
A Journal Published by Statistics Canada
Volume 38, Number 2, December 2012

Contents

Waksberg Invited Paper Series

Lars Lyberg	
Survey Quality	107

Regular Papers

Jaqueline Garcia-Yi and Ulrike Grote	
Data collection: Experiences and lessons learned by asking sensitive questions in a remote coca growing region in Peru	131
Jun Shao, Martin Klein and Jing Xu	
Imputation for nonmonotone nonresponse in the survey of industrial research and development	143
Jae Kwang Kim and Minsun Kim	
Riddles Some theory for propensity-score-adjustment estimators in survey sampling	157
Ian Plewis, Sosthenes Ketende and Lisa Calderwood	
Assessing the accuracy of response propensity models in longitudinal studies.....	167
Sarat C. Dass, Tapabrata Maiti, Hao Ren and Samiran Sinha	
Confidence interval estimation of small area parameters shrinking both means and variances.....	173
Dan Liao and Richard Valliant	
Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data.....	189
Qixuan Chen, Michael R. Elliott and Roderick J.A. Little	
Bayesian inference for finite population quantiles from unequal probability samples	203

Short Notes

Satkartar K. Kinney	
Multiple imputation with census data	215

Notice	219
Corrigendum.....	220
Acknowledgements.....	221
Announcements	223
In Other Journals.....	225

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.



Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work.

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2014 Waksberg Award.

This issue of *Survey Methodology* opens with the twelfth paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Elizabeth A. Martin (Chair), Mary E. Thompson, Steve Heeringa and J.N.K. Rao for having selected Lars Lyberg as the author of this year's Waksberg Award paper.

2012 Waksberg Invited Paper

Author: Lars Lyberg

Lars Lyberg, Ph.D., is former Head of the Research and Development Department at Statistics Sweden and currently Professor Emeritus at the Department of Statistics, Stockholm University. He is the founder of the Journal of Official Statistics (JOS) and served as Chief Editor for 25 years. He is chief editor of *Survey Measurement and Process Quality* (Wiley 1997) and co-editor of *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (Wiley 2010), *Telephone Survey Methodology* (Wiley 1988) and *Measurement Errors in Surveys* (Wiley 1991). He is co-author of *Introduction to Survey Quality* (Wiley 2003). He chaired the Leadership Group on Quality of the European Statistical System and chaired the Organizing Committee of the first European Conference on Quality in Official Statistics, Q2001. He is former president of IASS and former chair of the ASA Survey Methods Section. He is a fellow of the American Statistical Association and the Royal Statistical Society.

Survey Quality

Lars Lyberg¹

Abstract

Survey quality is a multi-faceted concept that originates from two different development paths. One path is the total survey error paradigm that rests on four pillars providing principles that guide survey design, survey implementation, survey evaluation, and survey data analysis. We should design surveys so that the mean squared error of an estimate is minimized given budget and other constraints. It is important to take all known error sources into account, to monitor major error sources during implementation, to periodically evaluate major error sources and combinations of these sources after the survey is completed, and to study the effects of errors on the survey analysis. In this context survey quality can be measured by the mean squared error and controlled by observations made during implementation and improved by evaluation studies. The paradigm has both strengths and weaknesses. One strength is that research can be defined by error sources and one weakness is that most total survey error assessments are incomplete in the sense that it is not possible to include the effects of all the error sources. The second path is influenced by ideas from the quality management sciences. These sciences concern business excellence in providing products and services with a focus on customers and competition from other providers. These ideas have had a great influence on many statistical organizations. One effect is the acceptance among data providers that product quality cannot be achieved without a sufficient underlying process quality and process quality cannot be achieved without a good organizational quality. These levels can be controlled and evaluated by service level agreements, customer surveys, paradata analysis using statistical process control, and organizational assessment using business excellence models or other sets of criteria. All levels can be improved by conducting improvement projects chosen by means of priority functions. The ultimate goal of improvement projects is that the processes involved should gradually approach a state where they are error-free. Of course, this might be an unattainable goal, albeit one to strive for. It is not realistic to hope for continuous measurements of the total survey error using the mean squared error. Instead one can hope that continuous quality improvement using management science ideas and statistical methods can minimize biases and other survey process problems so that the variance becomes an approximation of the mean squared error. If that can be achieved we have made the two development paths approximately coincide.

Key Words: Quality management; Total survey error; Quality framework; Mean squared error; Process variability; Statistical process control; Users of survey data.

1. Introduction

This article has been prepared in recognition of Joe Waksberg's unique contributions and leadership in survey methodology. My first encounter with Joe's work was his article on response errors in expenditure surveys written with John Neter (Neter and Waksberg 1964). Among other things that article introduced me to the cognitive phenomenon called telescoping. Later in life I had the opportunity to work with Joe on the first conference and monograph on telephone survey methodology where we were part of the editorial group (Groves, Biemer, Lyberg, Massey, Nicholls and Waksberg 1988). We also collaborated on the preparation of many of the Hansen Lectures that were published in the *Journal of Official Statistics* (JOS) during my term as its Chief Editor. Joe himself delivered the sixth lecture, which was published in JOS (Waksberg 1998). Joe was a fantastic leader and it is a great honor for me to have been invited to write this article on survey quality, a topic that occupied his mind a lot.

Many of my friends have conveyed their views or sent me materials in preparation of this article. Especially I want to thank Paul Biemer, Dan Kasprzyk, Fritz

Scheuren, Dennis Trewin, and Maria Bohata for helping me.

Survey quality is a vague, albeit intuitive, concept with many meanings. In this article I discuss some observations related to the development and treatment of the concept over the last 70 years and for some developments it is possible to trace roots that can be found even farther back. Most of my discussion, however, concerns current issues in government statistical organizations. It is within official statistics that most my survey quality examples take place.

The article is organized as follows: In Section 2 I discuss the total survey error paradigm, including error typologies, treatment of the errors, and survey design taking all error sources into account. In section 3 I discuss quality management philosophies that have had a large impact on survey organizations since the early 1990's. This impact is manifested by methods and approaches like recognition of the user or the client, a discussion of costs and risks in survey research, and the need for organizations to continuously improve. Section 4 provides examples of quality initiatives in survey organizations. Section 5 deals with the difficulties in measuring quality, either

1. Lars Lyberg, Department of Statistics, Stockholm University, 10691 Stockholm, Sweden. E-mail: Lars.Lyberg@stat.su.se.

directly or indirectly via indicators. How these measures should be communicated to the users or clients is also covered. Section 6, finally, offers some thoughts about how survey practices *must* change to better serve the needs of the users. The last section contains references.

2. The total survey error paradigm

2.1 Some history of survey sampling

There are a number of papers describing the development of early survey sampling methodology. In that early development there is an implicit or explicit recognition of quality issues although they are hidden under labels such as errors and survey usefulness (Deming 1944). The historical overviews provided by, for instance, Kish (1995), Fienberg and Tanur (1996), and O'Muircheartaigh (1997) all emphasize the fact that the period up to 1950 is characterized by a full-bloom development of sampling theory. During the 1920s the International Statistical Institute agreed to promote ideas on representative sampling suggested by Kiear (1897) and Bowley (1913). In 1934 Neyman published his landmark paper on the representative method. Later Fisher's (1935) randomization principle was used in agricultural sampling and Neyman (1938) developed cluster sampling, ratio estimation and two-phase sampling and introduced the concept of confidence interval. Neyman showed that the sampling error could actually be measured by calculating the variance of the estimator. Bill Cochran, Frank Yates, Ed Deming, Morris Hansen and many others further refined the concepts of sampling theory. Hansen led a research group at the U.S. Census Bureau where much of the applied work and new theory development was conducted in those days. One remarkable result of the Census Bureau efforts was the two-volume textbook on sampling theory and methods (Hansen, Hurwitz and Madow 1953). As a matter of fact the advances in sampling theory were so prominent at the time that Stephan (1948) found it worthwhile to write an article about the history of modern sampling methods.

It was early recognized that there could be survey errors other than those attributed to sampling. There are writings on the effects of question wording such as Muscio (1917). Research on questionnaire design was quite extensive in the 1940s. Problems with errors introduced by fieldworkers collecting agricultural data in India were addressed by Mahalanobis (1946), resulting in a method for estimating such errors. The method is called "interpenetration" and can be used to estimate, so called, correlated variances introduced by interviewers, editors, coders and those who supervise these groups. The most prominent error sources were certainly known around 1950. Deming had listed error sources (1944) that constitute the first published typology of

survey errors and Hansen and Hurwitz (1946) had discussed subsampling among nonrespondents in an attempt to provide unbiased estimates in a situation with an initial nonresponse. But the methodological emphasis, up to then, had been on developing sampling theory, which is quite understandable. It was very important to be able to show that surveys could be conducted on a sampling basis and in a variety of settings. By 1950 it had been demonstrated quite successfully that this was indeed possible. So it was time to move on to other issues and refinements.

In those early days the use of the word quality was confined to mainly quality control, sometimes as quality control of survey operations. It was common that the quality control was verification and/or estimation of error sizes for various operations. Statistics were known to be plagued by errors other than those stemming from sampling but the process quality issue of how to systematically reduce these errors and biases was still to be developed (Deming 1944; Hansen and Steinberg 1956).

The user 60 years ago was a somewhat obscure player, although not at all ignored by prominent survey methodology developers. For instance, Deming (1950) claimed that until the purpose is stated, there is no right or wrong way of going about a survey. Some other statisticians made similar statements. But the user was really hiding behind terms, such as subject-matter problem, study purpose or the key functions of a statistical system.

Even now survey and quality are vague concepts. As pointed out by Morganstein and Marker (1997) varying definitions of quality undermine improvement work so we should, at least, try to distinguish between different definitions to see what purposes they might serve. One of the most cited definitions is attributed to Joseph Juran, namely quality being a direct function of "fitness for use". It turns out that Deming already in 1944 used the phrase "fitness for purpose", not to define quality, but rather to explain what made a survey product work.

For a long time "good" quality was implicitly equivalent to a small mean squared error (MSE), *i.e.*, data should be accurate and accuracy of an estimate can be measured by MSE, which is the sum of the variance and the squared bias. We have noticed that survey statistics should also be useful, later denoted "relevant". Many of today's quality dimensions were not really an issue at the time. The users, too, were accustomed to the fact that surveys took time to carry out; timeliness was surely on the agenda but not as explicitly as it is today. A census took years to process. The users were accustomed to a technology that could only deliver relatively simple forms of accessibility. Hence, it was natural for users and producers to concentrate on making sure that the statistical problem coincided reasonably well with the subject-matter problem and that MSE was kept on a

decent level, where MSE many times was and still is equivalent with just the variance, without a squared bias term added.

Before proceeding any further, let us define “survey”. A *survey* is a statistical study designed to measure population characteristics so that population parameters can be estimated. Two examples of parameters are the proportion unemployed at a given time in a population of individuals, and the total revenue of a business or industry sector during a given time period. A survey can be defined as a list of prerequisites (Dalenius 1985a). According to Dalenius a study can be classified as a survey if the following prerequisites are satisfied:

1. The study concerns a set of objects comprising a population;
2. The population under study has one or more measurable properties;
3. The goal of the study is to describe the population by one or more parameters defined in terms of measurable properties, which requires observing (a sample of) the population;
4. To get observational access to the population a frame is needed;
5. A sample of objects is selected from the frame in accordance with a sampling design that specifies a probability mechanism and a sample size n (where n might equal N , the population size);
6. Observations are made on the sample in accordance with a measurement process (*i.e.*, a measurement method and a prescription as to its use);
7. Based on the measurements, an estimation process is applied to compute estimates of the parameters when making inference from the sample to the population under study.

This definition implicitly lists the specific error sources that are present in survey work. For each source there are a number of methods available that minimize the effects but also measure their sizes (Biemer and Lyberg 2003; Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau 2009).

Deviations from the definition reflect quality flaws. Moreover such deviations are common. In some designs selection probabilities are unknown or the variance estimator chosen might not be the most suitable one, given the sample design applied. Whether such flaws are problematic or not depends on the purpose.

2.2 The components of the total survey error paradigm

The total survey error paradigm is a theoretical framework for optimizing surveys by minimizing the accumulated size of all error sources, given budgetary constraints. In

practice this means that we want to minimize the mean squared error for selected survey estimates, namely those that are considered most important by the main stakeholders. The mean squared error is the most common metric for survey work consisting of a sum of variances and squared bias terms from each known error source. Groves and Lyberg (2010) provide a summary of the status of the paradigm in the past and in today’s survey practice.

The idea that surveys should be designed taking all error sources into account stems from the early giants in the field. Morris Hansen, Bill Hurwitz, Joe Waksberg, Leon Pritzker, Ed Deming and others at the U.S. Census Bureau, Leslie Kish at the University of Michigan, P.C. Mahalanobis at the Indian Statistical Institute, and Tore Dalenius, Stockholm University were among those who took the lead in survey research, emphasizing errors and optimal design. They worried about the inherent limitations associated with sampling theory since nonsampling errors could make the theory break down. They were very practical and thought a lot about balancing errors and the costs to deal with them. Some of them saw similarities between a factory assembly line (Deming and Geoffrey 1941) and the implementation of some of the survey processes and introduced control methods obtained from industrial applications.

Dalenius (1967) realized that there was as yet no “survey design formula” that could provide an optimal solution to the design problem. The approach taken by Dalenius and also Hansen, Hurwitz and Pritzker (1967) was a strategy of minimizing all biases and going for a minimum-variance scheme so that the variance became an approximation of the MSE. This was supposed to happen through intense verification schemes for ongoing productions and quite extensive evaluation studies for future productions. In 1969 Dalenius, inspired by Hansen, presented a paper on total survey design, where the word “total” reflected the thought about taking all error sources into account. Hansen, Hurwitz, Marks and Mauldin (1951), Hansen, Hurwitz and Bershad (1961), and Hansen, Hurwitz and Pritzker (1964) developed the U.S. Census Bureau Survey Model that reflected contributions from interviewers, coders, editors, and crewleaders and allowed the estimation of those contributions to the total survey error. These estimation schemes were elaborated on by Bailar and Dalenius (1969) and consisted of variations of replication and interpenetration. Bias estimation was assumed to be handled by comparing estimates obtained from the regular operations with those obtained from preferred procedures (that could not be used on a large scale due to financial, administrative or practical reasons). Today this kind of approach is called the “gold standard”.

It was stated that good survey design called for reasonably effective control of the total error by careful

specifications of the survey procedures, including adequate controls. Hansen, Deming and others did worry about control costs but although statistical process control and acceptance sampling had been implemented in a number of survey organizations, there was very little discussion about continuous process improvement. A lot of the quality work had to do with estimation of error rates, controlling error levels for individual operators and conducting large-scale evaluation studies that usually took a long time. Users were not directly involved in the design process but in the U.S. federal statistical system they had at least some influence on what should be collected and presented. Dalenius (1968) provides more than 200 references on users and user conferences associated with the products of the U.S. Federal statistical system.

While total survey design was first advocated by Hansen, Dalenius and others, users were seldom directly involved in the final determination of survey requirements. Quite often an official, administrator or statistician acted as a subject-matter specialist. Several decades ago this was the way we thought about users. Their opinions counted but they were not really involved in design decisions. Lurking in the back of our heads was the thought that this might not be a perfect model and in the late 1970's Statistics Sweden published an internal booklet called "What to do if a customer shows up on our doorstep".

The basic design approach suggested by Hansen, Dalenius and others contained a number of steps including:

- Specification of an ideal survey goal.
- Analysis of the survey situation regarding financial, methodological and information resources.
- Developing a small number of alternative designs.
- Evaluating the alternatives by reference to associated preliminary assessments of MSE and costs.
- Choosing one of the alternatives or a modification of one of them or deciding not to conduct a survey at all.
- Developing the administrative design including feasibility testing, a process signal system (currently called *paradata*), a design document, and a Plan B.

Kish (1965) had slightly different views on design. He liked the neo-Bayesian applications in survey sampling and psychometrics advocated by colleagues at the University of Michigan (Ericson 1969; Edwards, Lindman and Savage 1963). For instance, Kish liked the idea that judgment estimates of measurement biases might be combined with sampling variances to construct more realistic estimates of the total survey error. Regarding the optimization problem Kish thought that the multipurpose situation was economically favorable for surveys but that it could be difficult to decide on what to base the design on. If one principal

statistic can be identified then that alone can decide the design and if there are a small number of principal statistics a compromise design is possible but if statistics are too disparate a reasonable design might not exist. Kish also emphasized the need for design information obtained from pilot surveys and pretests to facilitate design decisions. Kish noted that survey design and measurement could vary greatly across environments while sampling did less so. That could be one reason that sampling can be easily placed among the traditional statistical theories and methods, while it is more difficult to place the survey process in one specific discipline (Frankel and King 1996 in their interview with Kish).

Kish, like the other giants, emphasized the importance of small biases but appreciated the fact that the reduction of one bias term might increase the total error. Kish was keen on getting a reasonable balance between different error sources and how error structures varied under different design alternatives. Like Hansen and colleagues Kish thought that relevant information should be contemporaneously recorded during implementation (again we see the parallel to *paradata*). Hansen and colleagues were really concerned about excessive but inadequate controls. They realized that some controls might have to be relaxed due to limited improvements and that degree of improvement in terms of affecting the estimates should be checked out before any relaxation could take place. They also suggested that one might have to compromise relevance to get controllable measurements or abstain from the survey. Both Hansen and colleagues and Kish were vigorously in favor of ending the practice that sampling error is the only survey error measured.

When we look at today's situation we can conclude that we still do not have a design formula for surveys. There is no planning manual to speak of and the literature on design is consequently very small, as is the literature on cost (Groves 1989 is an exception). And no design formula is in sight. Since the advent of the U.S. Census Bureau survey model a number of variants have appeared on the scene, some of them quite complicated (Groves and Lyberg 2010). A common characteristic is the fact that they tend to be incomplete, *i.e.*, they do not take all error sources into account. Most statistical attention is on variance components and especially on measurement error variance. There are a number of other weaknesses associated with the total survey error concept. Most notably a user perspective is missing and a vast majority of users are not in a position to question or even discuss accuracy. The complex error structures and interactions do not invite outside scrutiny and user contacts often tend to concern less technical issues such as timeliness, comparability and costs. Users are not really informed about real levels of accuracy and we know very

little about how users perceive information about errors and how to act on that.

As pointed out by Biemer (2001), in his discussion of Platek and Särndal (2001), there is a lack of routine measurements of MSE components in statistical organizations. There are good reasons for this state of affairs. Complexity has already been mentioned and to that we can add factors such as costs, the fact that it is almost impossible to publish such information at the time data are released, and that there is no measure of total error that would take all error sources into account, either because a lack of proper methodology or that some errors defy expression. Groves and Lyberg (2010) list some other weaknesses associated with the total survey error paradigm. For instance, we need to know more about the interplay between variances and biases. It is possible that an increase in simple response variance goes hand in hand with a reduction in response bias, say, when we compare interview mode with self-administrative alternatives. Recently, West and Olson (2010) showed that interviewer variance can occur not only from individual interviewers' effect on the responses within their assignments but also because individual interviewers successfully obtain cooperation from different groups of sample members.

Despite all its limitations, the strengths of the total survey error framework are quite convincing. The framework provides a taxonomic decomposition of errors, it separates variance from bias and observation from nonobservation, and it defines the different steps in the survey process. It serves as a conceptual foundation of the field of survey methodology, where subfields are defined by their associated error structures. Finally, it identifies the gaps in the research literature since any typology will show that some process steps are more "popular" than others. Just compare the respective sizes of the literatures on data collection and data processing.

It seems, however, as if the total survey error framework needs some expansion along lines some of which were identified half a century ago. We need some guidance on trade-offs between measuring error sizes and making processes more error-free. Spencer's (1985) question is: how much should we spend on measuring quality versus quality enhancement? We also need some guidance on how to integrate additional notions into the framework, so that it becomes a total survey quality framework rather than a total survey error framework (Biemer 2010). For instance, if "fitness for use" predominates as a conceptual base, how can we launch research that incorporates error variation associated with different uses? This aspect will be discussed in the next section.

3. Quality management philosophies in survey organizations

During the late 1980's and the early 1990's some statistical organizations were under severe financial pressure and in some cases simultaneously criticized for not being sufficiently attentive to user needs. Governments in Sweden, Australia, New Zealand and Canada as well as the Clinton administration in the U.S. were all keen on improving efficiency and user influence within their respective statistical systems. It was natural for these organizations to look for inspiration in management theories and methods (Drucker 1985) and specifically on what was called quality management (Juran and Gryna 1988). In that newer literature it was possible to study the role of the customer, leadership issues, the notion of continuous quality improvement, and various tools that could help the statistical organization improve. Especially influential to survey practitioners was work by Deming (1986), since he emphasized the role of statistics in quality improvement. He vigorously promoted the idea that improvement work should be led by statisticians, since they are trained in distinguishing between different kinds of process variation. He thought that there were too few statistical leaders advising top management in businesses and he wanted more proactive statisticians to become such leaders. He was especially keen on developing Shewhart's ideas about control charts as a means to distinguish between the different types of variation, namely common and special cause variation. Shewhart's improvement cycle Plan-Do-Check-Act was also part of Deming's thoughts on quality (Shewhart 1939).

Management principles have, of course, existed since ancient times. Juran (1995) provides lots of examples of what was in place in, for instance, the Roman empire. Craftsmanship and a guild system were basic building blocks. There were methods for choosing raw materials and suppliers. Processes were inspected and improved. Workers were trained and motivated and customers got warranties. All these features are found also in today's management systems. The more modern development includes quality frameworks or business excellence models such as Total Quality Management (TQM), International Organization for Standardization (ISO) standards, the Malcolm Baldrige quality award criteria, the European Foundation for Quality Management (EFQM), Six Sigma, Lean Six Sigma, and the Balanced Scorecard. These models are not totally different. They often share a common set of values and common criteria for excellence. Rather they represent a natural development that can be seen in all kinds of work.

Thus, there has been a gradual adoption of quality management models and quality strategies in statistical organizations and a merging with concepts and ideas already used in statistical organizations. My personal timeline for this development is the following (readers are invited to come up with different sets of events and dates):

1875	Taylor introduces what he called scientific management;
1900-1930	Taylor's ideas are used in, for instance, Ford's and Mercedes Benz's assembly lines;
1920's	Fisher starts developing theories and methods for experimental design;
1924	Shewhart develops the control chart;
1940	The U.S. War Department develops a guide for analyzing process data;
1944	Deming presents the first typology of survey errors;
1944	Dodge and Romig present theory and tables for acceptance sampling;
1946	Deming goes to Japan;
1950	Ishikawa suggests the fishbone diagram as a tool for identifying factors that have a profound effect on the process outcome;
1954	Juran goes to Japan;
1960	Many businesses embark on a zero defects program;
1960	The U.S. Census Bureau quality control programs are developed;
1961	The U.S. Census Bureau survey model is launched;
1965-1966	Kish and Slobodan Zarkovich start talking about data quality rather than survey errors;
1970's	Many statistical organizations provide quality guidelines;
1975	The Total Quality Management (TQM) framework is launched;
1976	The first quality framework in a statistical organization containing more dimensions than relevance and accuracy;
1987-1989	Launching of the ISO 9000, Malcolm Baldrige Award, Six Sigma and EFQM models;
1990's	Many statistical organizations start working with quality improvement and excellence models;
1997	The Monograph on Survey Measurement and Process Quality;
1998	Mick Couper introduces the concept "paradata" as a subset of process data;

2001	The Eurostat leadership group on quality organizes the first conference on Quality Management in Official Statistics;
2007	Business architecture ideas enter the survey world.

From the mid 1990's and on quality management philosophies have had an enormous effect on many statistical organizations. The effect is not necessarily higher quality across the board (no one has checked that). But the philosophies have led to an awareness in most organizations of the importance of good contacts with users and clients, and an aspiration in many of them to become "the best" or "world class". Quality is on the agenda.

3.1 The concept of quality

During the last decades it has become obvious that accuracy and relevance are necessary but not sufficient when assessing survey quality. Other dimensions are also important to the users. The development of survey quality frameworks has taken place mainly within official statistics and has been triggered by the rapid technology development and other developments in society. These advanced technologies have created opportunities and user demands regarding potential quality dimensions such as accessibility, timeliness, and coherence that simply were not emphasized before. Decision-making in society has become more complex and global resulting in demands for harmonized and comparable statistics. Thus, there is a need for quality frameworks that can accommodate all these demands. Several frameworks of quality have been developed and they each consist of a number of quality dimensions. Accuracy and relevance are just two of these dimensions.

For instance, the framework developed by OECD (2011) has eight dimensions: relevance, accuracy, timeliness, credibility, accessibility, interpretability, coherence, and cost-efficiency (Table 1). Similar frameworks have been developed by Statistics Canada (Statistics Canada 2002; Brackstone 1999), and Statistics Sweden (Felme, Lyberg and Olsson 1976; Rosén and Elvers 1999). The Federal Statistical System of the U.S. has a strong tradition in emphasizing the accuracy component (U.S. Federal Committee on Statistical Methodology 2001) although it certainly appreciates other dimensions. Perhaps they are viewed as dimensions of a more nonstatistical nature that still need a share of the total survey budget. The International Monetary Fund (IMF) has developed a framework that differs from those of OECD, Australian Bureau of Statistics, Statistics Sweden, and Statistics Canada. IMF's framework consists of a set of prerequisites and five dimensions of quality: integrity, methodological soundness,

accuracy and reliability, serviceability, and accessibility (see Weisman, Balyozov and Venter 2010).

Table 1
OECD's quality framework

Dimension	Description
Relevance	Statistics are relevant if users' needs are met.
Accuracy	Closeness between the value finally retained and the true, but unknown, population value.
Credibility	The degree of confidence that users place in data products based on their image of the data provider.
Timeliness	Time length between data availability and the event or phenomenon data describe.
Accessibility	How readily data can be located and accessed from within data holdings.
Interpretability	The ease with which the data user may understand and properly use and analyze the data.
Coherence	Reflects the degree to which data products are logically connected and mutually consistent.
Cost-efficiency	A measure of the costs and provider burden relative to the output.

Without sufficient accuracy, other dimensions are irrelevant but the opposite is also true. Very accurate data can be useless if they are released too late to affect important user decisions or if they are presented in ways that are difficult for the user to access or interpret. Furthermore, quality dimensions are often in conflict. Thus, providing a quality product is a balance act where informed users should be key players. Typical conflicts exist between timeliness and accuracy, since it takes time to get accurate data through, for instance, extensive nonresponse follow-up. Another conflict is the one between comparability and accuracy since application of new and more accurate methodology might disturb comparisons over time (Holt and Jones 1998).

Thus, many organizations have adopted a multi-faceted quality concept consisting not only of accuracy but also other dimensions. We might talk about a quality vector whose components vary slightly between organizations both in number and in contents. There are a number of problems associated with the quality vector approach.

First, the development has not been preceded by user contacts. Producers of statistics have believed that users are interested in a specific set of dimensions even though it is obvious that a vast majority of users think that error structures are too complicated to grasp and assume that the producer should be responsible for delivering the best possible accuracy. In cases where the user or client has specific accuracy requirements a more in-depth dialog can take place between the two. In the rare studies that have investigated user perceptions of information on quality it turns out that users are mostly interested in dimensions that are easily understood, such as timeliness and indicators that are seemingly straight forward, such as response rates. The

user wants the producing statistical organization to be credible, which translates into being capable of producing data with small or at least known errors and delivering them in a timely, reliable, and accessible fashion. The thought that it would be possible to produce a total quality measure based on weighted assessments of the different dimensions is not realistic, although Mirotchie (1993) argues to the contrary. In that paper Mirotchie makes a case for a standard set of quality indicators and provides a hypothetical illustration of indexing data quality indicators and computing an actual index (in this illustration the indicators are precision, nonresponse, reliability, timeliness and residuals). Even if a composite indicator in the form of an index were a possible development, the user would like to know which indicators contributed most to an index value. From a user's point of view the least favorable index value could still reflect a situation providing the highest quality. Rarely can a low accuracy be compensated by good ratings on other dimensions, not even in the case of election exit polls where timeliness is imperative. Accuracy is still necessary and there is wide agreement that all reputable organizations should meet accuracy standards (Scheuren 2001; Kalton 2001; Brackstone 2001). Phipps and Fricker (2011) provide an overview of quality frameworks and literature on total survey error. Thus, we can agree that survey quality is a multi-faceted concept involving multiple features of a statistical product or service.

3.2 The quality movement's impact on statistical organizations

Just extending the quality framework from one or two dimensions to several is not sufficient to create a quality environment. In the late 1980's and early 1990's many statistical organizations became interested in quality issues beyond traditional aspects of data quality. Issues concerning customer satisfaction, communicating with customers, competition, process variability, cost of poor quality, waste, business excellence models, core values, best practices, quality assurance, and continuous quality improvement were suddenly part of the everyday activities in many organizations.

Successful organizations know that continuous improvement (Kaizen) is necessary to stay in business and they have developed measures that help them change. This is true also for producers of statistics. Changes that are supposed to improve the statistical product are triggered by user demands, competition from other producers and from producer values that emphasize continuous improvement as part of the general business environment. The measures that can help a statistical organization improve are basically identical to those of other businesses. They can be built on business excellence models such as the European Foundation for

Quality Management (EFQM) (1999). The core values of the EFQM model include results orientation, customer focus, leadership and constancy of purpose, management by process measures and facts, personnel development and involvement, continuous learning, innovation and improvement, development of partnerships, and public responsibility. This model has been adopted by the European Statistical System (ESS) as a tool for national statistical institutes in Europe for achieving organizational quality. The thought is that good product quality, according to the dimensions mentioned (or some other product quality definition) cannot be achieved without good underlying processes used by the organization. It can also be argued that good product quality is achieved most efficiently and reliably by good process quality. If we view quality as a three-level concept it can be visualized as shown in Table 2.

3.2.1 Product quality

The deliverables agreed upon are called the product. It can be one or several estimates, datasets, analyses, registers, standard processes or other survey materials such as frames and questionnaires. Product quality is the traditional quality concept used when informing users or clients about the quality of the product or service. It can be measured and controlled by means of degree of adherence to specifications and requirements for product characteristics adding up to quality dimensions of a framework. Measures of accuracy and margins of error belong here. Also observations whether service levels agreements established with the client have been accomplished are relevant. In line with quality management principles, it is also quite common to conduct user satisfaction surveys to find out what users think about the products and services that are provided.

3.2.2 Process quality

All processes have to be designed so that they deliver what they are supposed to. This means that we have to have some kind of quality assurance perspective when processes are defined. For instance, the process of interviewing implies that a number of elements must be in place for the

process to deliver what is expected. Examples of elements are an effective selection of interviewers and a training program, a compensation system as well as supervision and feedback activities. Thus we aim at building quality into the process via the quality assurance. Quality control efforts are only used to check if the process works as intended. It cannot by itself be used to build quality into the process. In Section 4.4 this process view is discussed in more detail. Process quality is measured and controlled via selection, observation and analyses of key process variables, so called process data or paradata (Morganstein and Marker 1997; Couper 1998; Lyberg and Couper 2005). Theory and methods imported from statistical process control can help the producer distinguish between the two types of variation, common and special cause. As long as all variation is contained within the upper and lower control limits associated with the control charts chosen, the process is said to be in statistical control and no process improvements are really possible by trying to adjust individual outcomes. If there are observations falling outside of the control limits, usually set at 3 sigma, then we have indications of special cause variation that should be taken care of so that the variation after adjustment is brought back to common cause variation. The following P-chart illustrates a possible situation:

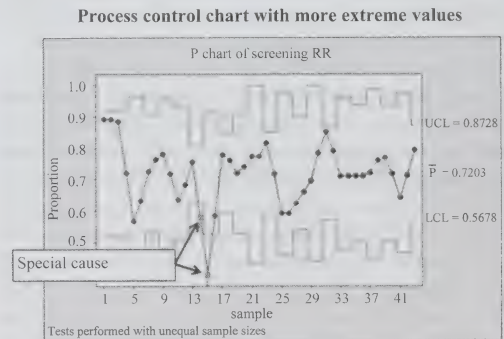


Table 2
Quality as a three-level concept*

Quality level	Main stake-holders	Control instrument	Measures and indicators
Product	User, client	Product specs, SLA, evaluation studies, frameworks, standards	Frameworks, compliance, MSE, user surveys
Process	Survey designer	SPC, charts, acceptance sampling, risk analysis, CBM, SOP, paradata, checklists, verification	Variation via control charts, other paradata analyses, outcomes of periodic evaluation studies
Organization	Agency, owner, society	Excellence models, ISO, CoP, reviews, audits, self-assessments	Scores, strong and weak points, user surveys, staff surveys

*SLA (Service Level Agreement), SPC (Statistical Process Control), CBM (Current Best Methods), SOP (Standard Operating Procedures), and COP (ESS Code of Practice).

Thus, the action sequence is the following. First the roots of the special causes are taken care of so that these variations are eliminated. After that the process displays common cause variation only. If that variation is deemed too large then the process has to change. The kinds of changes necessary are seldom obvious at the outset. Indeed perhaps several are necessary to decrease the process variation. Typically, a process improvement project is needed and the quality management literature has promoted a number of tools that are useful in such projects. Most of these tools are borrowed from statistics (control charts, experiments, regression analysis, Pareto diagrams, scatter plots, stratification) but there are also tools for identifying probable problem root causes (fishbone diagrams, process flow charts, brainstorming). The common thinking is that improvement projects should be “manned” by people working with the process or by people very much familiar with the process in other ways. Sometimes, we talk about forming an improvement team, where also the client or customer participates. In any improvement work suggested changes have to be tested. When Shewhart first developed his control charts he also suggested that improvement work should follow a sequence of operations, Plan-Do-Check-Act. What this sequence tells us is that any process changes suggested should be tested to see if they actually improve the process. If not, another change is made, and testing done again. Deming called this line of thinking the Shewhart cycle but since Deming spent a lot of time promoting it, many eventually called it the Deming cycle. The changes sought after could be decreased process variation, reduced costs, or increased customer satisfaction. The improvement project methodology is described in for instance Joiner (1994), Box and Friends (2006), Breyfogle (2003), and Deming (1986).

Another way of checking the process quality is to use acceptance sampling. Acceptance sampling (Schilling and Neubauer 2009) can be applied in situations where process elements can be grouped in batches. The batches are controlled and based on the outcome of that control it is decided whether the batch should be approved or reworked. Acceptance sampling plans guarantee an average outgoing quality in terms of, say, error rate, but there is no direct quality improvement involved. It is a control instrument that is suitable for operations such as coding, editing and scanning and then only when these processes are not really in statistical control. The method has been heavily criticized by Deming (1986) and others but can be the only control means available in situations where staff turnover is high and there is no time to wait for stable processes.

Global paradata (Scheuren 2001) are “error” rates of different kinds. Examples include nonresponse rates, coding error rates, scanning error rates, listing error rates, *etc.* In

some operations the error rates are calculated using verification, which means that the operation is repeated in some way. That is the case for the coding operation. In other operations the calculation can be based on a classification scheme, which is the case for nonresponse rate calculation. These global paradata tell us something about the process. They are process statistics, *i.e.*, summaries of data. A large nonresponse rate indicates problems with the data collection process and a high coding error rate indicates problems with the coding process. From these summaries it is sometimes possible to distinguish common and special cause variation and decide what action to take.

Some standardized processes can be controlled by means of simple checklists. Checklists are very effective when it is crucial that every process step is made and in the right order (Morganstein and Marker 1997). This is the case when airline pilots prepare for take-off. No matter how many times they have taken off, without a checklist the day will come when they forget an item. In statistics production sampling is such a process, albeit with less severe consequences if items are missed. It might very well be the case that a statistical organization has a standardized process for sample selection and a checklist that can be used as a combination of work instruction and control instrument.

There is a kind of checklist that can be used in more creative processes such as the overall survey design process. It is not possible to standardize the survey design process but it is possible to list a number of critical steps that always must be addressed. The list does not tell us how to address them. It just serves as a reminder that an individual step should not be omitted or forgotten. Morganstein and Marker (1997) discuss this kind of checklist and call them (and the simpler checklists) Current Best Methods (CBM). They describe the CBM development process and how the CBMs can be used to decrease the process variation in statistical organizations. For instance, an organization might have seven different imputation methods and systems in its toolbox. It is costly to maintain these seven systems. It is unlikely that they are equally efficient. If they are, it may not be economically feasible to keep them all. In this situation a CBM that describes fewer options to the organization seems like a good idea. This could be accomplished by forming an improvement team consisting of the imputation experts and some clients. CBMs are supposed to be revised when new knowledge is obtained, which implies that there is an expiration date associated with every CBM.

CBMs are of course “best practices” in some sense. Many organizations want best practices implemented and used. Morganstein and Marker offer a process for developing these best practices and keeping them current. It is beneficial for an organization if the variation in process

design can be kept at a minimum. It then becomes easier to train people and change the process when it becomes unstable or when new methods are developed. On the other hand, if CBMs and other standards are not vigorously enforced within an organization, they will not be widely used and the investment will not pay off.

3.2.3 Organizational quality

Management is responsible for quality in its widest sense. It is the organization that provides leadership, competence development, tools for good customer relations, investments, and funding. The quality management field has given us business excellence models that can help us evaluate our statistical organizations in the same way other businesses are evaluated. The two main business excellence models are the Baldrige National Quality Program and the European Foundation for Quality Management (EFQM).

These models consist of criteria to be checked when assessing an organization. The Malcolm Baldrige award uses seven main criteria: Leadership, strategic planning, customer and market focus, information and analysis, human resource focus, process management, and business results. Each criterion has a number of subcriteria. For instance, human resource focus consists of work systems, employee education, training and development, and employee well-being and satisfaction. The EFQM model has nine criteria: Leadership, strategy, people, partnerships & resources, processes, products & services, customer results, people results, society results, and key results. These models can be used for self-assessment or external assessment. The organization provides a description of what is in place regarding each criterion and the organization is scored based on that description. Typically self-assessments result in higher scores than external ones. It is very difficult to get a high score from external evaluators since the models are very demanding. For each criterion the organization is asked if it has a good approach in place somewhere in the organization. This is often the case. The next question is how wide-spread this good approach is within the organization. Many organizations lose momentum here, since there is very little truth in the mantra "the good examples are automatically spread throughout an organization". Instead good approaches usually have to be vigorously promoted before they are accepted within the organization. The third question asks whether the approach is periodically evaluated to check if it achieves the results expected. This is where most organizations fail. Their usual strategy is to exhaust an approach until the problems are so great that the approach has to be replaced rather than adjusted. This strategy is, of course, disruptive and expensive and does not score highly in excellence assessments. The maximum number of points

that can be obtained using these models is 1000 and very rarely does a winner get more than 450-600 points, which is an indication that there is a lot of room for improvement even in world class organizations.

Some statistical organizations have used business excellence models for assessment. The Czech Statistical Office was announced Czech National Quality Award Winner for 2009 in the Public Sector category based on EFQM. The office got 464 points. Eurostat's leadership group on quality recommended the European national statistical offices to use the EFQM as a model for their quality work and Finland and Sweden are among those that have done so. Since the leadership group released its report in 2001 (see Lyberg, Bergdahl, Blanc, Booleman, Grünwald, Haworth, Japac, Jones, Körner, Linden, Lundholm, Madaleno, Radermacher, Signore, Zilhao, Tzougas and van Brakel 2001) other frameworks and standards have been developed. The European Statistical System has launched its Code of Practice, which consists of a number of principles with associated indicators. Regarding some principles, however, the indicators are more like clarifications. The list of principles resembles other lists that have been developed by the UN and other organizations.

External assessments are probably more reliable than internal ones. There are a number of reasons for that. One is that it is difficult to criticize your peers since you have to interact with them in the future or if your own product or service will be assessed by those peers in the future. Experiences from Statistics Sweden and Statistics Canada show that self-assessments are limited in their capability of identifying serious weaknesses (see Section 5.3).

3.2.4 Some specific consequences for statistical organizations

Most statistical organizations have adopted quality management ideas to varying degrees and with varying success. As pointed out by Colledge and March (1993) it is possible to list a number of obstacles associated with such implementation. For a government agency it can be difficult to motivate its staff through monetary incentives, since there are restrictions on how tax money can be spent. The variety of users and products makes the dialog between the service provider and the user complicated and as mentioned neither the users, or for that matter the providers are totally familiar with all the biases and other quality problems that are present in statistics production. The effect of errors on the uses can vary and are often unknown. To complicate matters further, unlike most other businesses, suppliers are not very enthusiastic. In other businesses suppliers get paid while statistical organizations must motivate theirs, the respondents, who are seldom even given a cash incentive.

On the other hand statistical organizations have a great advantage when it comes to applying quality management principles. A statistical organization knows how to collect and analyse data that can guide improvement efforts. One of the cornerstones in quality management philosophies is that decisions should be based on data and businesses that do not have support from statisticians are often unaware of data quality problems, which can have consequences for their decision-making. By and large, though, a statistical organization is not different from any other business and it is quite possible to apply quality management ideas to improve all aspects of work.

4. Examples of quality initiatives in statistical organizations

In this section we will provide some examples of initiatives that statistical organizations have engaged in as a result of a general interest in quality in society.

4.1 The total survey error

Perhaps the most important thing to notice is that research and development in survey design, implementation, sampling and nonsampling errors, and the effect of errors on the data analysis continue to thrive. Data with small errors is the major goal for reputable organizations, which is indicated by the steady flow of textbooks on data collection, sampling, nonresponse, questionnaire design, measurement errors, and comparative studies. New textbooks are in progress covering gaps such as business surveys, translation of survey materials, and paradata. There are journals such as the *Journal of Official Statistics*, *Survey Methodology*, and *Survey Practice* that are entirely devoted to topics related to statistics production in a wide sense. Numerous other journals such as the *Public Opinion Quarterly*, the *Journal of the American Statistical Association*, and the *Journal of the Royal Statistical Society* devote much space to survey methods. The Wiley series on Survey Methodology and its associated conferences (on panel surveys, telephone survey methods (twice), measurement errors, process quality, business surveys, testing and evaluating questionnaires, computer assisted survey information collection, nonresponse, and comparative surveys) have been very successful and that is the case also for the continuing workshops on nonresponse and total survey error. Thus, there is no shortage of ideas regarding specific error sources and their treatment. Admittedly there are areas that are understudied such as specification errors, data processing errors and the impact of errors on the data analysis but by and large there is a healthy interest in knowing more about survey errors. The challenge lies in communicating this knowledge to people working in

statistical organizations and in developing design principles that can be used to improve statistics production. There is a noticeable gap between what is known through research and what is known and applied in the statistical organizations. Thus, staff capacity building seems to be a continuing need, especially since the common idea that good examples spread like ripples within and between organizations is a myth. If that indeed were the case quality would by now be fantastic everywhere. Since it is not, many organizations have developed extensive training programs (Lyberg 2002).

4.2 Risk and risk management

One element of quality management that has entered the survey world is risk and risk management. Eltinge (2011) even talks about Total Survey Risk as an alternative to the total survey error paradigm. The identification and management of risks is an important part of modern internal auditing (Moeller 2005) and is perhaps the only major element that is missing in quality management frameworks such as EFQM. An error source can be seen as more risky than another and should, therefore, be handled with more care and resources than another less risky. For instance, not having an effective system for statistical disclosure control is seen as a very risky situation. Unlawful data disclosure is very rare historically, but when it happens it could potentially destroy future data collection attempts. Certain design decisions can be seen as risky. For instance, if we choose a data collection method that does not fit the survey topic we might get estimates that are so far from the truth that the results are useless. An example might be to study sensitive behaviors using face to face or telephone interviewing instead of a self-administered mode. There are also technical risks that need to be identified and assessed. For instance, the U.S. National Agricultural Statistical Service (Gleaton 2011) like many others has plans for disaster recovery. Groves (2011) and Dillman (1996) both discuss how the production culture and the research culture within a statistical organization might view risks in different ways. Change in statistical organizations is generally slow and there are sometimes good reasons for that. Change might result in failures such as unsuccessful implementation, large costs and decreased comparability. So in some sense both producers and users have a tendency to be hesitant toward changes suggested by researchers and innovators and that might be one reason why change takes a long time. It is very common to have parallel measurements for some time to handle risks associated with implementing a new method or system. According to Groves (2011) the production culture and the users have had the final say about any changes, at least up until now. At the same time innovation is badly needed in many production systems and there are examples of stove-pipe organizations that do not have much time left

(to remain unchanged) because the resources to maintain their systems are simply not there. So even though there is resistance against change, lack of resources and competition will make sure that statistical organizations become more process-oriented and efficient. Reducing the number of systems and applications and developing and using more standardization seem to be one road forward.

4.3 The client/customer/user

The advent of quality management ideas in statistical organizations has made the receivers of statistical products and services more visible. Commercial firms have always talked about the client or the customer while government organizations have tended to call them users. In any case the recognition of someone who is supposed to use the endproducts has not been obvious to some providers. Admittedly the user has been a speaking partner since the beginning of the survey industry. In the U.S., conferences for users were quite frequent already 50 years ago (Dalenius 1968; Hansen and Voight 1967). During six months 1965-66, for example, the U.S. Census Bureau organized 23 user conferences across the country and there were also advisory groups. The advisory nature of contacts with users has prevailed in many countries. The user conference format still exists but user input is now complemented by other means such as public discussions and internet forums. Rarely have users been directly involved in the planning and design of surveys. Even when it comes to discussions about the quality of data, producers have acted as stand-in users. The quality frameworks are a good example. The quality dimensions were defined with minimal consultation with users. The literature on how users perceive information about quality is extremely limited (Groves and Lyberg 2010). Also, we do not know if the information on quality that we provide is useful to them (Dalenius 1985b). In fact, an educated guess is that many times it is not. In many surveys the users are many and sometimes unknown and their information and analytical needs cannot be foreseen ahead of time. It is often possible to single out one or a few main users to communicate with, but many of the design and quality problems are so complicated that a vast majority of users expect the service provider to deliver a product with the smallest possible error. Hansen and Voight stated that accuracy should be sufficient to avoid interpretation problems. Today there seems to be consensus among many that what users are interested in are products and services that can be trusted, *i.e.*, the service provider should be credible. It is impossible for most users to check levels of accuracy. Aspects that an average user can discuss are issues such as timeliness, accessibility and relevance. Detailed discussions about technical matters and design trade-off

issues including accuracy and comparability are more difficult to have.

During recent decades the user has indeed become more prominent. Some organizations develop service level agreements together with a main user or client, where requirements of the final product or service are listed and can be checked at the time of delivery. Many organizations conducting business surveys have created units that continuously communicate with the largest businesses, since their participation and provision of accurate information is absolutely essential for the estimation process (Willimack, Nichols and Sudman 2002). The large businesses are not users in the strict sense. They are important suppliers often with an interest in the survey results. Another common communication tool is the customer satisfaction survey. The value of such surveys is limited due to the acquiescence phenomenon and problems finding a knowledgeable respondent who is also willing to respond. Also, many customer satisfaction surveys are based on self-selection resulting in zero inferential value. In those surveys the results can only be viewed as lists of issues and concerns that some customers convey. Such information can, of course, be very valuable but is not suitable for estimation purposes. Many survey organizations now conduct user surveys on a continuing basis (Ecochard, Hahn and Junker 2008).

4.4 The process view

Quality management has reemphasized the importance of having a process view in statistics production. To view the production process as a series of actions or steps towards achieving a particular end that satisfies a user, leads to a good product quality. Process quality is an assessment of how far each step meets defined requirements or specifications. One way of controlling the process quality is to collect process data that can vary with each repetition of the process. The interesting process variables to monitor are those that have a large effect on the process's end result. Thus to check a process for stability and variation we need mechanisms for identifying, collecting and analysing these key process variables. The quality management science has given us tools such as the Ishikawa fishbone diagram to identify candidates for key process variables. The statistical process control methodology has given us tools to distinguish between special and common cause variation and how to handle these two variation types. Usually we use control charts originally developed by Shewhart (Deming 1986; Mudryk, Burgess and Xiao 1996) to make those distinctions. Then, again, we use methods from quality management to adjust the process if necessary. Examples include flowcharts, Pareto diagrams, and other simple means for the production team to identify the root causes of problems (Juran 1988).

Process data have been used to check on processes used in statistics production since the 1940's, first within the U.S. Census Bureau and then at Statistics Canada and to some extent also in other agencies. Typical processes that were checked included coding, keying and printing and the process data were mainly error rates. Some of the process checks used at the U.S. Census Bureau were so complicated and expensive that their value was questioned (Lyberg 1981), especially since the associated feedback loops were inefficient and not always aiming for the root causes of the errors. It was common that operators were blamed for system problems and at the time there was no emphasis on continuous quality improvement. The thinking at the time was more directed toward verification and correction.

Morganstein and Marker (1997) developed a generic plan for process continuous improvement that can be used in statistics production. They had worked in many statistical organizations since the 1980's and observed that quality thinking was not really developed in most of them. Their generic plan was built on their first-hand experiences and the general quality management ideas laid out by *e.g.*, Juran (1988), Deming (1986), Box (1990), and Scholtes, Joiner and Streibel (1996). In essence the plan consists of seven steps:

- The critical product characteristics are identified together with the user, both broad and more single effort needs.
- A map of the process flow is developed by a team familiar with the process. The map should include the sequence of process steps, decision points and customers for each step.
- The key process variables are identified among a larger set of process variables.
- The measurement capability is evaluated. It is important that decisions are based on good data, not just data. Available data might be useless. This is an area where statistical organizations should have an advantage over other organizations. One should not reach conclusions about process stability without knowledge about measurement errors. Above all, data should allow quantification of improvement.
- The stability of the process is determined. The variability pattern of the process data is analyzed using control charts and other statistical tools.
- The system capability is determined. If stability is not achieved after special cause variation has been eliminated an improvement effort is called for. System changes must be made when the process variation is so large that it does not meet specifications, such as minimum error rates or production deadlines. Typical methods to reduce variation are the development and implementation of a new training program or the

enforcement of a standard operating procedure. The latter can be a process standard, a current best methods standard or a simple checklist.

- The final step of the improvement plan is to establish a system for continuous monitoring of the process. We cannot expect processes to remain stable over time. For many reasons they usually start drifting after some time. A monitoring system helps keeping track of new error structures, new customer requirements, and the potential of improved methods and technology and can suggest process improvements.

The Morganstein and Marker book chapter had a distinct effect on quality work and process thinking in many European statistical organizations. Interest in these issues increased and some organizations started their own quality management system where process improvement was central.

At the 1998 Joint Statistical Meetings Mick Couper presented an invited paper on measuring quality in a CASIC environment. He meant that the new technology generated lots of by-product data that could be used to improve the data collection process. He named those paradata, not in his paper but in his session presentation. This naming caught on very quickly in the survey community and it made sense to define the trilogy data, metadata, and paradata. Thus we had one term for data about the data (metadata) and another for data about the process (paradata). Obviously paradata are process data but for a long time paradata were confined to data about the data collection process, while the term used in many European statistical organizations was "process data" and took all survey processes into account (Aitken, Hörngren, Jones, Lewis and Zilhao 2004). Recently a renewed broadening of the meaning of the concept has taken place. Kennickell, Mulrow and Scheuren (2009) remind us about what they call macro paradata, global process data such as response rates, coverage rates, edit failure rates, and coding error rates that always have been indicators of process quality in statistical organizations. Lyberg and Couper (2005), Kreuter, Couper and Lyberg (2010), and Smith (2011) also use the more inclusive meaning of paradata where other processes than data collection are taken into account. There is a risk that paradata, like quality, becomes an overused concept. There are examples of discussions where all data, apart from the survey estimates, are considered paradata, which, of course, does not make sense.

Paradata is a great naming and they are necessary to judge process quality. However, a word of caution is in place. One should never collect paradata that are not related to process quality and it is important to know how to analyze them. Sometimes statistical process control methods

can be used but at other times other analytical techniques are needed. For instance, to be able to control interviewer falsification we might need to look at several processes simultaneously, but theory and methodology for such analysis might not be readily available.

The expanded use of microdata that concern individual records, such as keystroke data and flagged imputed records, is an effect of using new technology. Modern data collection procedures generate enormous amounts of these kinds of paradata but so do systems for computer-assisted manual coding and systems for pure automated coding as well as systems for scanning of data. It makes no sense to confine the concept to data collection.

Quality management has taught us to prevent process problems rather than fix them when they appear, that it is important to distinguish between different types of process variation since they require different actions, that any process intervention or improvement should be based on good data and proper analysis methods, and that even stable processes eventually start drifting, which calls for continuous monitoring.

4.5 Standardization and similar tools

One way of keeping process quality in control is to reduce variation by encouraging the use of standards and similar documents. Colledge and March (1997) discuss four classes of documents.

- A standard is a document that should be adhered to almost without exception. Deviations are not recommended and require approval of senior management. Corrective action should be taken when a standard is not fully met. An organization can become certified according to a standard. This is the case for ISO standards, where a few are relevant to statistical organizations.
- A policy should be applied without exceptions. For instance, an organization can have a policy regarding the use of incentives to boost response rates.
- Several organizations have developed guidelines for different aspects of the statistics production. Typically, guidelines can be skipped if there are “good” reasons to do so.
- A recommended practice is promoted but adherence is not mandatory.

Admittedly, the categories of this classification scheme are not mutually exclusive, especially if we also take language and cultural aspects into account. For instance, in the Swedish language policies and guidelines are very close conceptually. If we consult the unauthorized but consensus based Wikipedia it says that “policies describe standards while guidelines outline best practices for following these

guidelines”. This sentence contains three of the categories mentioned by Colledge and March. It is probably best to relate to these different kinds of documents in a similar fashion. They all attempt to improve quality by reducing various types of variation and we should not dwell too much on what they are called.

Although standards have been an important part of survey methodology for a long time they have gained momentum since statistical organizations became interested in quality management. Early standards such as Hansen *et al.* (1967) and U.S. Bureau of the Census (1974) concentrated on discussing the presentation of errors in data. At the U.S. Census Bureau all publications should inform users that data were subject to error, that analysis could be affected by those errors, and that estimated sampling errors are smaller than the total errors. For major surveys the nonsampling errors should be treated in more detail unlike in the past. Many other statistical organizations imported this line of thinking. For instance, the quality frameworks mentioned earlier are expansions including also other quality dimensions than accuracy. The European Statistical System has successively developed and launched what was first called Model Quality Reports and currently just Standard for Quality Reports (Eurostat 2009a). The standard provides recommendations to European National Institutes (notice the conceptual complexity) for preparation of quality reports for a “full” range of statistical processes and their outputs. The standard treats the basic quality dimensions relevance, accuracy, timeliness, accessibility, coherence and comparability.

Let us look at some examples. Regarding measurement error, which is part of the accuracy component, the standard says that the following information should be included in a quality report:

- Identification and general assessment of the main risks in terms of measurement error.
- If available, assessments based on comparisons with external data, reinterviews or experiments.
- Information on failure rates during data editing.
- The efforts made in questionnaire design and testing, information on interviewer training and other work on error reduction.
- Questionnaires used should be annexed in some form.

Regarding timeliness the standard says that the following information should be included:

- For annual or more frequent releases: the average production time for each release of data.
- For annual and more frequent releases: the percentage of releases delivered on time, based on scheduled release dates.
- The reasons for nonpunctual releases explained.

There are also sections on how to communicate information regarding trade-offs between quality dimensions, assessment of user needs and perceptions, performance and cost, respondent burden as well as confidentiality, transparency and security. Even though there is a section on user needs and perceptions, users have obviously not been involved in the preparation of the standard itself. We still know very little about how users perceive and use information about quality. The standard is backed by a much more detailed handbook for quality reports (Eurostat 2009b) and both documents are built around the 15 principles listed in the European Statistics Code of Practice, which is the basic quality framework for the European Statistical System. The Code of Practice principles concern professional independence, mandate for data collection, adequacy of resources, quality commitment, statistical confidentiality, impartiality and objectivity, sound methodology, appropriate statistical procedures, nonexcessive burden on respondents, cost-effectiveness, relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, and, finally, accessibility and clarity. Each principle is accompanied by a set of indicators that the individual organization can measure to establish whether it meets the Code or not. Some indicators are vague and very subjective in nature such as “the scope, detail and cost of statistics are commensurate with needs”, while others are more specific, such as “a standard daily time for the release of statistics is made public”. Peer reviews of compliance to a limited set of the principles have been conducted using an earlier version of the Code and, not surprisingly, many national statistical offices in Europe have problems living up to the Code (Eurostat 2011a). Therefore in order to assist the implementation of the Code a supporting framework has been developed, called the Quality Assurance Framework (QAF) that contains more specific guidance regarding methods and references (Eurostat 2011b). This seems to be a very useful document since its references are mainly summaries of the state-of-the-art in areas such as sampling, questionnaire design, editing and so on, which stimulates conformity to current best practices.

The Code of Practice has many similarities with the UN Fundamental Principles of Official Statistics (de Vries 1999). The latter promotes also the principle of international cooperation and coordination, which is, to a large extent, an element that is missing in today’s development of statistics production (Kotz 2005). Even neighbouring countries can have very different approaches and methodological competence levels and the differences are sometimes difficult to explain. Experience shows that global development collaboration is difficult to achieve. We meet, we talk, and we bring back ideas that might fit our own systems. It is harder to agree on common approaches. One global standard that

relates to statistics production is the ISO 20252 on market, opinion and social research (International Standards Organization 2006). This is a process standard with around 500 requirements concerning the research activities within an organization. It is a minimum standard for what to do rather than how to do things. It is suitable for organizations that conduct surveys and the organization can apply for certification. In April 2010 more than 300 organizations world-wide had been certified, most of them marketing firms. One national statistical office (Uruguay) was certified in 2009 and Statistics Sweden is planning a certification in 2013 but those are the only national offices that have chosen this path. The standard concerns the organization’s system for quality management, management of the executive elements of the research, data collection, data management and processing, and reporting on research projects (Blyth 2012).

The standards of the U.S. Federal Statistical System concentrate on the accuracy component. Although not formally a standard the U.S. Federal Committee on Statistical Methodology (2001) suggests various methods for measuring and reporting sources of error in surveys. In 2002 the U.S. Office of Management and Budget (OMB) issued information quality guidelines (OMB 2002) whose purpose was to ensure and maximize the quality, objectivity, utility, and integrity of information disseminated by federal agencies. OMB (2006a) has also issued standards and guidelines for surveys. They are built in a standard fashion. First comes a standard such as “Response rates must be computed using standard formulas to measure the proportion of the eligible sample that is represented by the responding units in each study, as an indicator of potential nonresponse bias”. This standard is then followed by a number of guidelines on how to make the necessary calculations while the final guideline states that “If the overall nonresponse rate exceeds 20%, an analysis of the nonresponse bias should be conducted to see whether data are missing completely at random”. As in the case of the ESS standards, the OMB guidelines are complemented by a supporting document (OMB 2006b) that can facilitate adherence to the standards.

Most agencies in the decentralized U.S. Federal Statistical System have documents in place that adapt the OMB guidelines. For instance, the U.S. Census Bureau has its own statistical quality standards that goes into more technical detail compared to the OMB documents. Each standard is described via requirements and sub-requirements and they often provide very specific examples of studies that can be conducted. Examples of other U.S. agencies that have standards related to the quality of information disseminated include the National Center for Health Statistics, National Center for Education Statistics, and the

Energy information Administration. All these standards can be downloaded from the agencies' websites.

Statistics Canada has issued quality guidelines since 1985. They are similar to the ESS guidelines since not just accuracy is emphasized. But they are much more detailed and contain lots of references. A special feature is that for some processes the guidelines prescribe the use of statistical process control. No other agency seems to be doing that. The latest edition of the guidelines is provided in Statistics Canada (2009).

Many other statistical organizations in the world have their own quality standards. They are sometimes described as guidelines or standards and sometimes as business support systems or quality assurance frameworks. In any case, the contents and style vary across organizations but the variation should be manageable. It should be possible to achieve higher degrees of standardization globally, since that has happened in other fields, such as air travel. Apted, Carruthers, Lee, Oehm and Yu (2011) discuss various ways to industrialize the statistical production process at the Australian Bureau of Statistics.

The question is whether international standards would benefit survey quality in general. Some areas where standards would be beneficial include computation of frequently used quality indicators such as error rates and design effects, as well as best practices for translation of survey materials, handling non-native language respondents, and weighting for nonresponse. One must bear in mind that once a standard is issued it has to be continually updated and it is well-known that they can be difficult to enforce. If they are comprehensive, standards can overwhelm the practitioner and, as a result, unless mandated and audited, they are largely ignored.

4.6 Statistical business process models

During recent years concepts like business process models and business architecture have become part of quality work in some statistical organizations. To make production processes more efficient and flexible they can be seen as part of a business architecture model (Reedman and Julien 2010). In statistics production a generic statistical process model is jointly developed by UNECE, Eurostat, and OECD. Any system redesign should be driven by customer demands, risk assessments and new developments. The architectural principles behind this thinking are summarized in Doherty (2010), which discusses architecture renewal at Statistics Canada.

Some of the principles are:

- Decision-making should be corporately optimal, which entails centralization of informatics, methodology support and processing.
- Use of corporate services such as collection, data capture and dissemination should be optimized.
- Reuse should be maximized by having the smallest possible number of distinct business processes and the smallest possible number of computer systems.
- The corporate toolkit should be minimized.
- There should be staff proficiency in tools and systems.
- Rework such as repeated editing should be eliminated.
- The focus should be on the core business and the work with support processes should be outsourced.
- Development should be separated from the on-going operations.
- Electronic data collection should be viewed as the initial mode.
- Structural obstacles, such as overlapping or unclear mandates should be removed.

These principles are very similar to those we identify when we apply quality management principles from the various frameworks and excellence models described previously. The principles represent a move from decentralization to more corporate level thinking. Many statistical organizations realize that stove-pipe thinking is a thing of the past and that a move to more centralization is necessary.

5. Measuring quality

Thus, quality is a multi-faceted concept and measuring it is a complicated task. We have noted that survey quality can be viewed as a three-dimensional concept associated with the final product, the underlying processes that lead to the product, and the organization that provides the means to carry out the processes and deliver the product or service in a successful way. There are basically two ways to measure quality. One is to directly estimate the total survey error or some components thereof. The other is to measure indicators of quality with the hope that they indeed reflect the concept itself.

5.1 Direct estimates of the total survey error

The existing decompositions of the mean squared error described in, for instance, Hansen *et al.* (1964), Fellegi (1964), Anderson, Kasper and Frankel (1979), Biemer and Lyberg (2003), Weisberg (2005), and Groves *et al.* (2009) are all incomplete in the sense that they do not reflect all error sources. It is seldom possible to compute the MSE directly in practical survey situations because this usually requires a parameter estimate that is essentially error free. However, it is possible to obtain a second best estimate of the true parameter value if there are resources available to collect data using some "gold standard" methodology that is not affordable or practical in a normal survey setting. This is

the standard evaluation methodology when the true parameter value can be uniquely defined. Gold standard methods are seldom error-free but they can to varying extents provide better estimates, and the difference between the regular estimate and the gold standard estimate can serve as an estimate of the bias, which is the methodology used in census post enumeration surveys (United Nations 2010). Often an evaluation concerns a specific error component such as census undercount, nonresponse bias, interviewer variance or simple response variance, since we want information not on total survey error per se but rather on the components' relative contribution to the total survey error so that root causes of problems can be identified and relevant processes improved. Large evaluation studies are very rare since they are so demanding and their value is sometimes questioned (United Nations 2010). Smaller regular evaluation studies, on the other hand, are necessary to get indications of process and methodological problems.

5.2 Indicators of quality

Continuing reporting of total survey error is a formidable task and no survey organization does that. Instead organizations provide indicators or statements regarding quality. For instance, according to Eurostat's (2009a) handbook for quality reports the following indicators should be measured:

- Coefficient of variation;
- Overcoverage rate;
- Edit failure rate;
- Unit response rate;
- Item response rates;
- Imputation rates;
- Number of mistakes;
- Average size of revisions.

The common theme here is that these paradata summary items are indicators that can be calculated without conducting special studies. The set of indicators that can be calculated directly from the survey data is by definition quite limited and their value questionable. For instance, to include overcoverage but not undercoverage just because only the former can be calculated directly from the available data does not make sense. It is undercoverage that poses the greatest coverage problem in surveys. Admittedly, the handbook prescribes the producer to assess the potential for bias (both sign and magnitude) but it is not clear how this should be accomplished. The producer is urged to include evaluation and quality control results, if such information exists as well. Level of effort measures for processes such as questionnaire design and coder training would be welcomed. There is no standard reporting format for such qualitative and quantitative information. In any case, the key

indicator list becomes severely limited when compared to the full list of main error sources and it is hard to see how they are perceived by the users and how they can be used by the producer to improve the process.

The producer needs a more complete list of indicators to be able to measure or assess various levels of quality to make sure that the design implementation is in control or to be able to mount a quality improvement project. The initial survey design must be modified or adapted during the implementation to control costs and maximize quality. Biemer (2010) discusses four strategies for reducing costs and errors in real time, *i.e.*, continuous quality improvement (CQI), responsive design (Groves and Heeringa 2006), Six Sigma (Breyfogle 2003), and adaptive total design and implementation.

When the continuous quality improvement strategy is used, key process variables are identified and so are process characteristics that are critical to quality (CTQ). For each CTQ, real-time, reliable metrics for the cost and quality are developed. The metrics are continuously monitored during the process and intervention is done to ensure that costs and quality are within acceptable limits. The responsive design strategy was developed to reduce nonresponse bias in face to face interviewing. It includes three phases. In the experimental phase a few design options are tested (*e.g.*, regarding incentive level). In the main data collection phase the option chosen in the experimental phase is implemented and the implementation continues until phase capacity is reached. In the nonresponse follow-up phase special methods are implemented to reduce nonresponse bias and control the data collection costs. Such methods include the Hansen-Hurwitz double sampling scheme, increased incentives, and using more experienced interviewers. Again the efforts continue until further reductions of the nonresponse bias are no longer cost-effective. Six Sigma is the most developed business excellence model since it relies so heavily on statistical methods. It contains a large set of techniques and tools that can be used to control and improve processes. Adaptive total design and implementation combines control features of CQI, responsive design and Six Sigma and does that so that it simultaneously monitors multiple error sources. Biemer and Lyberg (2012) give several examples of CTQs and metrics for various survey processes. For instance, regarding the measurement process attributes that are CTQs might include the abilities to identify and repair problematic survey questions, to detect and control response errors, and to minimize interviewer biases and variances. Corresponding metrics might include missing data item by question, refusal rate by size of business, results of replicate measurements, suspicious edits actually changed, and field work results by interviewer. The metrics can be analyzed using statistical process control or

analysis-of-variance methodologies. Different related metrics can be displayed together in a dashboard fashion. For instance if one CTQ is the ability to discover interviewer cheating we might want to have a dashboard showing the metrics average interview length by interviewer and the distribution of some sensitive sample characteristic, also by interviewer.

5.3 Self-assessments and audits

The quality management philosophy has introduced the concepts of self-assessment and audit into statistics production. We are anxious to know what users, clients, owners and other stakeholders think about the products and services provided by the statistical organization. There are a number of tools available for this kind of evaluation. We have already mentioned the customer satisfaction survey. Other tools include employee surveys, internal audits and external audits. Customer surveys can shed light on what users think about products and services provided. They can be used to determine user needs and to identify what product characteristics really matter to the users. Another line of questioning might concern the image of the organization and how it compares to the images of other organizations, be they competitors or not. The customer satisfaction survey is very common in society. Often it cannot be used to make inference to the target population of users due to its methodological and conceptual shortcomings. The abundance of satisfaction surveys in society, developed and implemented by people with no formal training in survey methods, contributes to lukewarm receptions in more serious settings resulting in nonresponse and measurement errors. For instance, the 2007 Eurostat User Satisfaction Survey consisted of two separate surveys. One was launched on the Eurostat webpage and the target population consisted of 3,800 registered users. Only those registered users that entered the website during the data collection period were exposed to the survey request and this led to a response rate around 5%. The second survey used email that was sent to a number of main users identified by Eurostat. This more controlled environment generated a response rate of 28%. These surveys also have problems identifying the most suitable respondent. If the “wrong” respondent is chosen within an organization this will most certainly lead to uninformed and misleading results.

The simplest type of self-assessment is the questionnaire or checklist that is filled out by the survey manager. An example is one from Statistics New Zealand. It is a checklist that consists of a number of indicators or assertions such as “information needs are regularly assessed through user consultation”, “good and accessible documentation”,

“indicators of accuracy regularly produced and monitored”, and “presentation standards met”. The manager is asked to answer yes or no to each assertion and make a comment if deemed necessary. Statistics Sweden had a similar system in place where one of the questions was “has overall quality of your product improved, declined or stayed the same compared to last year?” When results were compiled for these three categories for the entire organization, a very small proportion of the managers reported declining quality, a somewhat larger proportion reported improved quality, while a vast proportion reported status quo. The managers simply did not have the proper means to assess overall quality. Furthermore, vague quantifiers like “regularly”, “good”, and “meeting standards” invite generous assessments. Also most managers do not want to look bad and status quo becomes a perfect escape route. This system of self-assessment was eventually abandoned by Statistics Sweden. It is possible to increase the value of these assessments by asking additional questions concerning details about how and when quality work was conducted. Some organizations use internal teams that audit important products. Julien and Royce (2007) describe a quality audit of nine products at Statistics Canada, where the purposes were to identify weaknesses and their root causes as well as identifying best practices. Review teams of assistant managers were formed so that each reviewer reviewed three different programs. The main weakness with an approach like this is the internal feature itself. Every reviewer knows that sooner or later it is his or her turn to be reviewed and there is a risk that this fact might hold them back. It is also internal in the sense that users are not explicitly present in the review process. In its general audit program on data quality management, however, Statistics Canada puts great emphasis on its user liaison system (Julien and Born 2006), which is one of the five systems forming the agency’s quality assurance framework, the others being corporate planning, methods and standards, dissemination, and program reporting.

A further variant of self-assessment is when it precedes an external audit. Statistics Netherlands (1997) describes how the Department of Statistical Methods is assessed by its staff. The assessment resulted in a listing of weak and strong areas that were later examined by an external team. Typically an external audit uses some kind of benchmark like a set of rules, a standard, or a code of practice for assessment purposes. The audit then results in a number of recommendations for the organization or the individual product or service.

Recently a general system for evaluating the total survey error has been developed at Statistics Sweden. Sweden’s Ministry of Finance wants quality evaluation results to be able to monitor quality improvements over time. Survey

quality must be assessed for many surveys, administrative registers, and other programs within the agency so there is need for some indicators that can serve as proxies for actual measures of quality. At the same time, the assessment process must be thorough, the reporting simple and the results credible. For each of the error sources specification, frame, nonresponse, measurement, data processing, sampling, model/estimation, and revision eight key products were rated poor, fair, good, very good, and excellent regarding each of five criteria. The criteria were knowledge of risks, communication with users, compliance with standards and best practices, available expertise, and achievement toward risk mitigation and/or improvement plans. The rating guidelines varied by criterion. For knowledge of risks they were:

An Example of the rating guidelines – Knowledge of risks

Poor ●	Fair ▲	Good ○	Very Good ☾	Excellent ⊙
Internal program documentation does not acknowledge the source of error as a potential factor for product accuracy.	Internal program documentation acknowledges error source as a potential factor in data quality.	Some work has been done to assess the potential impact of the error source on data quality.	Studies have estimated relevant bias and variance components associated with the error source and are well-documented.	There is an ongoing program of research to evaluate all the relevant MSE components associated with the error source and their implications for data analysis. The program is well-designed and appropriately focused, and provides the information required to address the risks from this error source.
	But: No or very little work has been done to assess these risks	But: Evaluations have only considered proxy measures (for example, error rates) of the impact with no evaluations of MSE components	But: Studies have not explored the implications of the errors on various types of data analysis including subgroup, trend, and multivariate analyses	

The evaluation process started with a self-assessment done by each of the eight key products. These reports and other relevant documents were studied by two external reviewers who then met with product owners and their staff to discuss the product processes. After that the reviewers presented detailed assessments and scored each product. The procedure identified important areas to improve within but also across products. In this first evaluation round measurement error turned out to be a problematic area for

almost all the key products. As any other approach at measuring or indicating total survey error this one does not really reflect total mean squared error. It requires thorough documentation of processes and improvements made and it is highly dependent on the skills and knowledge of the external reviewers. This study is reported in Biemer, Trewin, Japec, Bergdahl and Pettersson (2012).

5.4 Quality profiles

In continuing surveys there is an opportunity to develop quality profiles. Such documents contain all that is known about the quality of a continuing survey or other statistical product assembled over a number of years. Quality profiles exist for only a few major surveys, all, except one, conducted in the U.S., including the Current Population Survey (Brooks and Bailer 1978), the Survey of Income and Program Participation (Jabine, King and Petroni 1990; Kalton, Winglee and Jabine 1998), the Schools and Staffing Survey (Kalton, Winglee, Krawchuk and Levine 2000), and the American Housing Survey (Chakrabarty and Torres 1996). The exception is the British Household Panel Survey (Lynn 2003). The main problem with a quality profile is that it is not timely, since it compiles results from often time-consuming studies of quality. The goal of the quality profile is to identify areas where knowledge about errors is deficient so that improvements can be made. Kasprzyk and Kalton (2001) and Doyle and Clark (2001) review the use of quality profiles in the U.S.

6. Where do we go from here?

Quality management ideas have been influential in many survey organizations. Concepts such as leadership, quality culture, problem prevention, customer, competition, risk assessment, process thinking, improvement, business excellence, and business architecture are increasingly discussed by leaders of survey organizations, e.g., Trewin (2001), Pink (2010), Fellegi (1996), Brackstone (1999), de Vries (1999), Groves (2011), and Bohata (2011). It seems as if the survey community is moving in a direction where statistics production becomes more streamlined and cost-effective but the pace is slow. Some organizations have started using a quality management model for self-assessment and steering purposes. EFQM is the recommended model for national statistical institutes within the European Statistical System and a couple of institutes, the Czech Republic and Finland, have even applied for their respective national EFQM awards. Some marketing firms are certified according to the ISO 9001 quality management standard and others are certified according to the ISO 20252 standard for market, opinion, and social research. This development ought to result in quality improvements but we cannot be really sure

until we start collecting relevant data. One thing is sure, though. Some customers prefer service providers that are certified, have won awards or can show evidence that they are working according to some quality framework or model. Very few customers would think that this is a negative thing.

The margins of error that we associate with estimates are usually too short, since they do not include all sources of variation. Point estimates can be off due to biases. Ideally it would be good if we were able to produce estimates of the total survey error instead of what we produce today. Such a development is, however, not realistic. We are not in a position to produce such estimates, not even occasionally, for reasons that have to do with finances, timing and methodology. That leaves us with indicators of total survey error and its components. Such indicators are of limited value to the users. Users simply do not know what to do with information on nonresponse rates, response variance measured by reinterviews or edit failure rates. On the other hand, such indicators are very useful to the producers of surveys. For instance, reinterview studies can identify fabrication and survey questions with poor response consistency. A majority of users appreciate the service provider's credibility and part of the credibility is the ability to present accurate data. Another important part of credibility is the willingness of the providers to evaluate their own quality and to report the results of such evaluations. Even if these evaluations show problems, it is better for the provider to find the problems than if entities outside the provider's organization find them. Most users do not want to become involved in discussions about errors and trade-offs between errors and for good reasons. It is simply too technical and confusing. If we accept that a good process quality is a prerequisite for a good product quality, we should gradually improve the processes so that they approach ideal bias-free ones. In that way the variance of an estimate becomes a good approximation of the mean squared error.

Despite endless discussions and a myriad of survey quality initiatives, practices have not changed much (Lynn 2004; Pink, Borowik and Lee 2010; Groves 2011; Bohata 2011). Perhaps the lack of competence within survey organizations is one root cause of the slow pace. Many theories and methodologies including statistics, IT, management, communication, and behavioral sciences are needed in survey research. The behavioral sciences are needed to identify the root causes of nonsampling errors. If errors are just quantified no improvement can happen. Current training programs emphasize sampling, non-response, coverage and estimation in the presence of these. Other processes and error sources such as measurement and data processing are not dealt with to the same extent. This

leads to a situation where studies on measurement error and data processing error are rare compared to studies on, say, nonresponse. There is a considerable confusion regarding concepts and methods in both the producer and the user camps. Another cause of slow pace might be the consensus philosophy that rules in some organizations when it comes to decision-making regarding changes. This philosophy is one of compromise. Input from many stakeholders is gathered and a decision is usually based on the smallest common denominator, which is never a good standard. Furthermore, arriving at this compromise usually takes a long time and lots of resources. This approach is very far from Plan-Do-Check-Act.

Survey quality is not an absolute entity. Current quality reporting a la one-size-fits-all is not working since fitness for use is defined by each user. Quality dimensions such as timeliness, comparability and accessibility should be decided together with main users while best possible accuracy given various constraints is the responsibility of the service provider.

Have the survey quality discussion and the adoption of quality management strategies resulted in better data? We do not know. Survey quality has not been assessed in a before-after fashion. There is a tendency towards greater standardization and centralization, which should prove cost-efficient but when it comes to data quality some indicators point in the wrong direction. For instance, in many countries nonresponse rates are increasing and error properties of mixed-mode, translation of survey materials, and other design features are not fully known or are different across cultures. There is no design formula, which results in shaky trade-off decisions and problems deciding about intensities with which quality control should be applied. There is a persistent quest for best practices in survey organizations but implementation is difficult and scattered. There is definitely a great need for an upgrade in the competence level across the board. A structured international competence development program for service providers is necessary as is a systematic international collaboration on how to best design and implement surveys. We must serve our users better by providing data with small errors. We can do this by better combining our knowledge about statistics and cognitive phenomena with the principles of quality management. The great positive note is the overwhelming positive attitude toward quality improvement among statistical organizations around the world.

References

- Aitken, A., Hörngren, J., Jones, N., Lewis, D. and Zilhao, M. (2004). *Handbook on improving quality by analysis of process variables*. Office for National Statistics, UK.

- Anderson, R., Kasper, J. and Frankel, F. (1979). *Total Survey Error: Applications to Improve Health Surveys*. San Francisco: Jossey-Bass.
- Apted, L., Carruthers, P., Lee, G., Oehm, D. and Yu, F. (2011). Industrialisation of statistical processes, methods and technologies. Paper presented at the International Statistical Institute Meeting, Dublin.
- Bailar, B., and Dalenius, T. (1969). Estimating the response variance components of the U.S. Bureau of the Census' Survey Model. *Sankhyā*, B, 341-360.
- Biemer, P. (2001). Comment on Platek and Särndal. *Journal of Official Statistics*, 17(1), 25-32.
- Biemer, P. (2010). Overview of design issues: Total survey error. In *Handbook of Survey Research*, (Eds., P. Marsden and J. Wright), Second Edition. Emerald Group Publishing Limited.
- Biemer, P., and Lyberg, L. (2003). *Introduction to Survey Quality*. New York: John Wiley & Sons, Inc.
- Biemer, P., and Lyberg, L. (2012). Short course on Total Survey Error. The Joint Program in Survey Methodology (JPSM), April 16-17, Washington, DC.
- Biemer, P., Trewin, D., Japac, L., Bergdahl, H. and Pettersson, Å. (2012). A tool for managing product quality. Paper presented at the Q Conference, Athens.
- Blyth, B. (2012). ISO 20252: Turning frameworks into best practice. Paper presented at the Q Conference, Athens.
- Bohata, M. (2011). Fit-for-purpose statistics for evidence based policy making. Memo, Eurostat.
- Bowley, A.L. (1913). Working-class households in reading. *Journal of the Royal Statistical Society*, 76(7), 672-701.
- Box, G. (1990). Good quality costs less? How come? *Quality Engineering*, 3, 1, 85-90.
- Box, G., and Friends (2006). *Improving Almost Anything: Ideas and Essays*. New-York: John Wiley & Sons, Inc.
- Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25, 2, 139-149.
- Brackstone, G. (2001). How important is accuracy? *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.
- Breyfogle, F. (2003). *Implementing Six Sigma*. Second Edition. New-York: John Wiley & Sons, Inc.
- Brooks, C., and Bailar, B. (1978). An error profile: Employment as measured by the Current Population Survey. Working paper 3, Office of Management and Budget, Washington, DC.
- Chakrabarty, R., and Torres, G. (1996). American Housing Survey: A Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.
- Colledge, M., and March, M. (1993). Quality management: Development of a framework for a statistical agency. *Journal of Business and Economic Statistics*, 11, 157-165.
- Colledge, M., and March, M. (1997). Quality policies, standards, guidelines, and recommended practices. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin), New-York: John Wiley & Sons, Inc.
- Couper, M. (1998). Measuring Survey Quality in a CASIC Environment. Paper presented at the Joint Statistical Meetings, American Statistical Association, Dallas, TX.
- Dalenius, T. (1967). Nonsampling Errors in Census and Sample Surveys. Report No. 5 in the research project Errors in Surveys. Stockholm University.
- Dalenius, T.E. (1968). Official statistics and their uses. *Review of the International Statistical Institute*, 26(2), 121-140.
- Dalenius, T. (1969). Designing descriptive sample surveys. In *New Developments in Survey Sampling*, (Eds., N.L. Johnson and H. Smith), New-York: John Wiley & Sons, Inc.
- Dalenius, T. (1985a). *Elements of Survey Sampling*. Swedish Agency for Research Cooperation with Developing Countries. Stockholm, Sweden.
- Dalenius, T. (1985b). Relevant official statistics. *Journal of Official Statistics*, 1(1), 21-33.
- Deming, E. (1944). On errors in surveys. *American Sociological Review*, 9, 359-369.
- Deming, E. (1950). *Some Theory of Sampling*. New-York: John Wiley & Sons, Inc.
- Deming, E. (1986). *Out of the Crisis*. MIT.
- Deming, W.E., and Geoffrey, L. (1941). On sample inspection in the processing of census returns. *Journal of the American Statistical Association*, 36, 215, 351-360.
- De Vries, W. (1999). Are we measuring up...? Questions on the performance of national systems. *International Statistical Review*, 67, 1, 63-77.
- Dillman, D. (1996). Why innovation is difficult in government surveys (with discussions). *Journal of Official Statistics*, 12, 2, 113-198.
- Doherty, K. (2010). How business architecture renewal is changing IT at Statistics Canada. Paper presented at the Meeting on the Management of Statistical Information Systems. Daejeon, South Korea, April 26-29.
- Doyle, P., and Clark, C. (2001). Quality profiles and data users. Paper presented at the International Conference on Quality in Official Statistics (Q), Stockholm.
- Drucker, P. (1985). *Management*. Harper Colophone.
- Ecochard, P., Hahn, M. and Junker, C. (2008). User satisfaction surveys in Eurostat and in the European Statistical System. Paper presented at the Q conference, Rome, Italy.
- Edwards, W., Lindman, H. and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Eltinge, J. (2011). Aggregate and systemic components of risk in total survey error models. Paper presented at ITSEW 2011, Quebec, Canada.
- Ericson, W. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 195-233.

- European Foundation for Quality Management (1999). *The EFQM Excellence Model*. Van Haren.
- Eurostat (2009a). ESS Standard for Quality Reports. Eurostat.
- Eurostat (2009b). ESS handbook for Quality Reports. Eurostat.
- Eurostat (2011a). European statistics Code of Practice. Eurostat.
- Eurostat (2011b). Quality assurance framework (QAF). Eurostat.
- Fellegi, I. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- Fellegi, I. (1996). Characteristics of an effective statistical system. *International Statistical Review*, 64, 2, 165-197.
- Felme, S., Lyberg, L. and Olsson, L. (1976). *Kvalitetsskydd av data*. (Protecting Data Quality.) Liber (in Swedish).
- Fienberg, S., and Tanur, J. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review*, 64, 237-253.
- Fisher, R. (1935). *The Design of Experiments*. New York: Hafner.
- Frankel, M., and King, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 1, 65-87.
- Gleaton, E. (2011). Centralizing LAN services. Memo, National Agricultural Statistics Service, U.S. Department of Agriculture.
- Groves, R. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.
- Groves, R. (2011). The structure and activities of the U.S. Federal Statistical System: History and recurrent challenges. *The Annals of the American Academy of Political and Social Science*, 631, 163, Sage.
- Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls, W. and Waksberg, J. (Eds.) (1988). *Telephone Survey Methodology*. New-York: John Wiley & Sons, Inc.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2009). *Survey Methodology*, Second Edition. New-York: John Wiley & Sons, Inc.
- Groves, R., and Heeringa, S. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, A*, 169, 439-457.
- Groves, R., and Lyberg, L. (2010). Total survey error: Past, present and future. *Public Opinion Quarterly*, 74, 5, 849-879.
- Hansen, M., and Hurwitz, W. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 517-529.
- Hansen, M., Hurwitz, W. and Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 32nd Session, 38, Part 2, 359-374.
- Hansen, M., Hurwitz, W. and Madow, W. (1953). *Sample Survey Methods and Theory*. Volumes I and II. New-York: John Wiley & Sons, Inc.
- Hansen, M., Hurwitz, W., Marks, E. and Mauldin, P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190.
- Hansen, M., Hurwitz, W. and Pritzker, L. (1964). The estimation and interpretation of gross differences and simple response variance. In *Contributions to Statistics*, (Ed., C. Rao). Oxford: Pergamon Press, 111-136.
- Hansen, M., Hurwitz, W. and Pritzker, L. (1967). Standardization of procedures for the evaluation of data: Measurement errors and statistical standards in the Bureau of the Census. Paper presented at the 36th session of the International Statistical Institute.
- Hansen, M., and Steinberg, J. (1956). Control of errors in surveys. *Biometrics*, 462-474.
- Hansen, M., and Voigt, R. (1967). Program guidance through the evaluation of uses of official Statistics in the United States Bureau of the Census. Paper presented at the International Statistical institute meeting, Canberra, Australia.
- Holt, T., and Jones, T. (1998). Quality work and conflicting policy objectives. *Proceedings of the 84th DGINS Conference*, May 28-29, Stockholm, Sweden. Eurostat.
- International Standards Organization (2006). Market, Opinion and Social Research. ISO Standard No. 20252.
- Jabine, T., King, K. and Petroni, R. (1990). Survey of Income and Program Participation (SIPP): Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.
- Joiner, B. (1994). *Generation Management*. McGraw-Hill.
- Julien, C., and Born, A. (2006). Quality management assessment at Statistics Canada. *Proceedings of the Q Conference*, Cardiff, UK.
- Julien, C., and Royce, D. (2007). Quality review of key indicators at Statistics Canada. *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*, 1113-1120.
- Juran, J.M. (1988). *Juran on Planning for Quality*. New York: Free Press.
- Juran, J.M. (1995). *A History of Managing for Quality*. ASQC Quality Press.
- Juran, J., and Gryna, F. (Eds.) (1988). *Juran's Quality Control Handbook*, 4th Edition. McGraw-Hill.
- Kalton, G. (2001). How important is accuracy? *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.
- Kalton, G., Winglee, M. and Jabine, T. (1998). *SIPP Quality Profile*. U.S. Bureau of the Census, 3rd Edition.
- Kalton, G., Winglee, M., Krawchuk, S. and Levine, D. (2000). *Quality Profile for SASS Rounds 1-3: 1987-1995*. Washington, DC: U.S. Department of Education.
- Kasprzyk, D., and Kalton, G. (2001). Quality profiles in U.S. Statistical Agencies. *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm 14-15 May 2001, CD-ROM.

- Kennickell, A., Mulrow, E. and Scheuren, F. (2009). Paradata or process modeling for inference. Paper presented at the Conference on Modernization of Statistics Production, Stockholm, Sweden.
- Kiear, A.N. (1897). The representative method of statistical surveys. *Kristiania Videnskaps-selskabets Skrifter: Historik-filosofiske Klasse*, (in Norwegian), 4, 37-56.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Kish, L. (1995). *The Hundred Years' Wars of Survey Sampling*. Centennial Representative Sampling, Rome.
- Kotz, S. (2005). Reflections on early history of official statistics and a modest proposal for global coordination. *Journal of Official Statistics*, 21, 2, 139-144.
- Kreuter, F., Couper, M. and Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Lyberg, L. (1981). *Control of the Coding Operation in Statistical Investigations: Some Contributions*. Ph.D. dissertation, Stockholm University.
- Lyberg, L. (2002). Training of survey statisticians in government agencies-A review. Invited paper presented at the Joint Statistical Meetings, American Statistical Association, New-York.
- Lyberg, L., Bergdahl, M., Blanc, M., Booleman, M., Grünewald, W., Haworth, M., Japac, L., Jones, L., Körner, T., Linden, H., Lundholm, G., Madaleno, M., Rademacher, W., Signore, M., Zilhao, M.J., Tzougas, I. and van Brakel, R. (2001). Summary report from the Leadership Group (LEG) on Quality. Eurostat.
- Lyberg, L., and Couper, M. (2005). The use of paradata in survey research. Invited paper, International Statistical Institute, Sydney, Australia.
- Lynn, P. (Ed.) (2003). *Quality Profile: British Household Panel Survey: Waves 1 to 10: 1991-2000*. Colchester: Institute for Social and Economic Research.
- Lynn, P. (2004). Editorial: Measuring and communicating survey quality. *Journal of the Royal Statistical Society*, Series A, 167.
- Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.
- Mirotichie, M. (1993). Data quality: A quest for standard indicators. *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 729-734.
- Moeller, R. (2005). *Brink's Modern Internal Auditing*. Sixth Edition. New-York: John Wiley & Sons, Inc.
- Morganstein, D., and Marker, D. (1997). Continuous quality improvement in statistical agencies. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 475-500.
- Mudryk, W., Burgess, M.J. and Xiao, P. (1996). Quality control of CATI operations in Statistics Canada, Memo, Statistics Canada.
- Muscio, B. (1917). The influence of the form of a question. *The British Journal of Psychology*, 8, 351-389.
- Neter, J., and Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 305, 18-55.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- Neyman, J. (1938). *Lectures and Conferences on Mathematical Statistics and Probability*. U.S. Department of Agriculture, Washington, DC.
- OECD (2011). Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities. OECD.
- O'Muirheartaigh, C. (1997). Measurement errors in surveys: A historical perspective. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 1-25.
- Phipps, P., and Fricker, S. (2011). Quality measures. Memo, Office of Survey Methods Research, U.S. Bureau of Labor Statistics.
- Pink, B., Borowik, J. and Lee, G. (2010). The case for an international statistical innovation program-Transforming national and international statistics systems. Paper presented at the Collaboration Leaders Workshop, April 19-23, Sydney, Australia.
- Platek, R., and Särndal, C.-E. (2001). Can a statistician deliver? *Journal of Official Statistics*, 17, 1, 1-20 and Discussion, 21-27.
- Reedman, L., and Julien, C. (2010). Current and future applications of the generic statistical business process model at Statistics Canada. Paper presented at the Q Conference, Helsinki.
- Rosén, B., and Elvers, E. (1999). Quality concept for official statistics. *Encyclopedia of Statistical Sciences*, New-York: John Wiley & Sons, Inc., update Volume 3, 621-629.
- Scheuren, F. (2001). How important is accuracy? *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.
- Schilling, E., and Neubauer, D. (2009). *Acceptance Sampling in Quality Control*, 2nd Ed. Chapman and Hall/CRC.
- Scholtes, P., Joiner, B. and Streibel, B. (1996). *The Team Handbook*. Joiner Associates Inc.
- Shewhart, W.A. (1939). *Statistical Methods from the Viewpoint of Quality Control*. U.S. Department of Agriculture, Washington, DC, U.S.A.
- Smith, T. (2011). Report on the International Workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. NORC/University of Chicago.
- Spencer, B. (1985). Optimal data quality. *Journal of the American Statistical Association*, 80, 564-573.
- Statistics Canada (2002). Statistics Canada's Quality Assurance Framework, Catalogue No.12-586-XIE, Ottawa.

- Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth Edition, Ottawa.
- Statistics Netherlands (1997). A self assessment of the Department of Statistical Methods. Research paper No. 9747, Statistics Netherlands.
- Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.
- Trewin, D. (2001). The importance of a quality culture. *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.
- United Nations (2010). *Post Enumeration Surveys: Operational Guidelines*. Department of Economic and Social Affairs, Statistics Division.
- U.S. Bureau of the Census (1974). *Standards for Discussion and Presentation of Errors in Data*. U.S. Department of Commerce, Bureau of the Census.
- U.S. Federal Committee on Statistical Methodology (2001). *Measuring and Reporting Sources of Errors in Surveys*, Statistical Policy Working Paper 31, Washington, DC: U.S. Office of Management and Budget.
- U.S. Office of Management and Budget (2002). Guidelines for ensuring, and maximizing the quality, objectivity, utility, and integrity of information disseminated by Federal agencies. Federal register, 67, 36, February 22.
- U.S. Office of Management and Budget (2006a). *Standards and Guidelines for Statistical Surveys*. U.S. Office for Management and Budget.
- U.S. Office of Management and Budget (2006b). Questions and answers when designing surveys for information collection. U.S. Office for management and Budget.
- Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 2, 119-137.
- Weisberg, H. (2005). *The Total Survey Error Approach*. The University of Chicago Press.
- Weisman, E., Balyozov, Z. and Venter, L. (2010). IMF's data quality assessment framework. Paper presented at the Conference on Data Quality for International Organizations, Helsinki, May 6-7.
- West, B., and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 5, 1004-1026.
- Willimack, D., Nichols, E. and Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. In *Survey Nonresponse*, (Eds., R. Groves, D. Dillman, J. Eltinge and R. Little), 213-228.
- Zarkovich, S. (1966). *Quality of Statistical Data*. Food and Agricultural Organization of the United Nations: Rome, Italy.

Data collection: Experiences and lessons learned by asking sensitive questions in a remote coca growing region in Peru

Jaqueline Garcia-Yi and Ulrike Grote¹

Abstract

Coca is a native bush from the Amazon rainforest from which cocaine, an illegal alkaloid, is extracted. Asking farmers about the extent of their coca cultivation areas is considered a sensitive question in remote coca growing regions in Peru. As a consequence, farmers tend not to participate in surveys, do not respond to the sensitive question(s), or underreport their individual coca cultivation areas. There is a political and policy concern in accurately and reliably measuring coca growing areas, therefore survey methodologists need to determine how to encourage response and truthful reporting of sensitive questions related to coca growing. Specific survey strategies applied in our case study included establishment of trust with farmers, confidentiality assurance, matching interviewer-respondent characteristics, changing the format of the sensitive question(s), and non enforcement of absolute isolation of respondents during the survey. The survey results were validated using satellite data. They suggest that farmers tend to underreport their coca areas to 35 to 40% of their true extent.

Key Words: Coca; Cocaine; Sensitive question; Misreporting; Nonresponse; Peru.

1. Introduction

Over the last 30 years, surveys have been increasingly used to explore sensitive topics (Tourangeau and Yan 2007). For example, data obtained from surveys have been used to investigate “socially undesirable” behaviors, such as the prevalence of illicit drug use (*e.g.*, Botvin, Griffin, Diaz, Scheier, Williams and Epstein 2000; Fergusson, Boden and Horwood 2008), illegal abortion (*e.g.*, Johnson-Hanks 2002; Varkey, Balakrishna, Prasad, Abraham and Joseph 2000), or alcohol consumption among adolescents (*e.g.*, Strunin 2001; Zufferey, Michaud, Jeannin, Berchtold, Chossis, van Melle and Suris 2007). Such surveys have been commonly utilized in academic research and policy analysis (Davis, Thake, and Vilhena 2009), even though asking sensitive questions has generally been seen as problematic. The responses have been considered to be prone to error and bias because respondents consistently underreport socially undesirable behaviors (Barnett 1998; Tourangeau and Yan 2007). Low response rates have been an additional concern. Those who are selected for a survey can simply refuse to take part in the survey or they can participate but refuse to answer the sensitive questions (Tourangeau and Yan 2007).

Recent surveys at the household level have incorporated sensitive questions related to the extent of coca growing areas (see *e.g.*, Ibanez and Carlsson 2010). Coca is a native bush from the Amazon rainforest in South America from the leaves of which cocaine is extracted. Colombia’s coca bush area represents 40%, Peru’s 40%, and Bolivia’s 20% of the total area under coca cultivation worldwide, amounting to

154,100 hectares (UNODC 2011). In Peru and Bolivia, the leaves of this bush have been traditionally used for many purposes from around 3000 B.C. (Rivera, Aufderheide, Cartmell, Torres and Langsjoen 2005) until today. Those traditional uses mainly include coca chewing and coca tea drinking to overcome fatigue, hunger and thirst; and to relieve “altitude sickness” and stomach ache symptoms, respectively (Rospigliosi 2004). Since the 1970s, however, coca cultivation skyrocketed because of its use as the raw material for the production of cocaine (Caulkins, Reuter, Iguchi and Chiesa 2005). The cocaine content of the coca leaves is below 1%, and ranges from 0.13 to 0.86% (Holmstedt, Jaatmaa, Leander and Plowman 1977). Therefore narcotics traffickers need large quantities of coca leaves to obtain enough of the alkaloid for commercialization in the illegal market. In general, growing coca for the narcotics trafficking business is a profitable activity. In fact, the income of a coca growing farmer has been calculated to be 54% higher than the income of a non coca growing farmer (Davalos, Bejarano and Correa 2008).

Consequently, coca-related research has become oriented towards evaluating the profitability of coca versus other cash crops (see, *e.g.*, Gibson and Godoy 1993; Torrico, Pohlen and Janssens 2005). Different attempts were made to replace coca by other crops, but it has been generally established that crop substitution as an anti-drug policy has been a failure (UNODC 2001). Decision makers and researchers have recognized that there are relevant socio-economic determinants that lead to coca growing other than economic profitability. These include social capital (Thoumi 2003),

1. Jaqueline Garcia-Yi, chair of Agricultural and Food Economics Technical University of Munich Weihenstephaner Steig 22, 85350, Freising, Germany. E-mail: jaqueline.garcia-yi@um.de; Ulrike Grote, Professor, Institute for Environmental Economics and World Trade, Leibniz University Hannover, Königsworther Platz 1, 30167 Hannover, Germany. E-mail: grote@iuw.uni-hannover.de.

saving account functions and financial reserve for large expenses (Bedoya 2003; Mansfield 2006). Comprehensive databases which include specific household-level information for coca growing areas are required to test those latter hypotheses.

Coca growing is not illegal *per se* in Peru (During the 1990s, the primary focus of the Peruvian Government was on “pacifying” the country by bringing terrorist groups under control. The Peruvian Government implemented what is currently known as the “Fujimori Doctrine”. The idea underlying this Doctrine was that the coca cultivation was not criminal in nature, but attributable to poverty. Consequently, the Fujimori Doctrine decriminalized all coca farmers, which diminished the farmers’ need for protection from terrorist associations, therefore making it easier for the Government to fight those violent groups (Obando 2006).), which partly reflects the social acceptance of traditional uses of coca in this country (UNODC 2001). Thus, the current legal framework seems to facilitate narcotics trafficking because coca used in illegal trade can be cultivated under the guise of traditional uses (INCB 2009; Durand 2005). Accordingly, Garcia and Antezana (2009) suggest that some farmers sell coca to those who purport to be traditional-use traders, but are actually narcotics traffickers who process coca leaves in different places, such as small towns at the border with Bolivia.

Even though coca farming is not illegal, coca-growing regions which are perceived to be supplying narcotics traffickers (e.g., regions with large coca fields) can be targeted by the Government for the implementation of forced eradication programs (Obando 2006). After eradication, coca growers are likely to incur large economic losses, depending on the total extent of their individual coca cultivation areas. Thus, some of the farmers might be reluctant to provide information on whether or not they have any coca under cultivation. It should also be expected that some of the farmers who admit to cultivating coca, would not report the true extent of the area, given their fear that large coca fields could be more prone to eradication.

Since there are both political and policy concerns in accurately and reliably measuring coca growing areas, it is necessary for survey methodologists to determine how to encourage response and truthful reporting of answers to sensitive questions related to coca growing. This article suggests and evaluates a number of strategies to increase both the reporting and the reliability of household-level responses in a remote coca growing region in Peru.

Although the topic of this article is specifically related to coca growing, the lessons learned about survey design and implementation could be used as a reference for dealing with other sensitive topics such as health-related issues (e.g., anti-conception and sexual behavior) or undesirable

behaviors (e.g., illegal drug use) in other regions in different countries.

The structure of the article is as follows: Section 2 describes the community in Peru subject to study, the specific strategies to reduce non-response and misreporting as well as the lessons learned from data collection related to sensitive questions in the research area. Section 3 presents the coca growing-related survey results and their validation, while Section 4 is comprised of a summary of the main results followed by the conclusion.

2. Data collection in a coca-growing community in rural Peru

This section describes the coca-growing community, and the primary data collection strategies applied in our study and the lessons learned.

2.1 Description of the research area

The research area was located in the Upper Tambopata valley at the border with Bolivia, one of the most remote and difficult to access Amazon rainforest areas in Peru (UNODC Office in Peru 1999). This valley lies in the Vilcabamba-Amboro Biodiversity Corridor in close proximity to national protected areas (see Figure 1). The entire population of the upper Tambopata valley is composed of immigrants, especially descendants from the Aymara indigenous population. Aymara is a native ethnic group originally from the Andes and Altiplano regions of South America. During the 1950s, most of the farmers were seasonal immigrants who left their Altiplano subsistence plots for only three to six months every year, and made the 320 km journey to the upper Tambopata valley to cultivate coffee on their individually owned agricultural plots (Collins 1984). Over time, most farmers became permanent settlers in the upper Tambopata valley, and cultivate coffee as their main cash crop (*ibid*).

Before 1989, coca cultivation in the upper Tambopata valley was very minor. Small-scale coca production was limited to self-consumption or local markets for traditional uses such as coca chewing by Andean farmers and miners. After 1989, coca cultivation was intensified, primarily in the neighboring upper Inambari valley. The change did not appear to be in response to increases in local demand or external demand by traditional users (UNODC Office in Peru 1999). Coca from those valleys is considered as low quality due to its bitterness, and it is in less demand for traditional chewing than coca from Cuzco region (Caballero, Dietz, Taboada and Anduaga 1998). Those increases were therefore related to narcotic traffic demand. In recent years, large increases in coca cultivation in the upper Tambopata valley have been consistently reported by the

United Nations (UN), as observed in Table 1. The percentage variation per year in the upper Tambopata valley is above the annual change of around 4% at national level.

Table 1
Coca cultivation in the upper Tambopata Valley (2005-2008)*

Year	Hectares	Percentage of variation in relation to previous year
2005	253	-
2006	377	49.0
2007	863	128.9
2008	940	8.9

*Since 2009 coca areas from the upper Tambopata valley are aggregated with coca areas from Inambari valley in UNODC reports. Therefore, it is not possible to estimate the percentage of variation in relation to previous year only for Tambopata valley during later years.

Source: Own calculation using data from UNODC (2009).

Coca provided by the upper Tambopata valley and upper Inambari valley seems to mainly supply cross border trade associations between Peruvian and Bolivian narcotics traffickers. Bolivia remains the world's third largest producer of cocaine, and it is a significant transit zone for cocaine of Peruvian-origin (U.S. Department of State 2009). Those valleys constitute a strategic coca production area for narcotics traffickers due to their proximity to an external exit route (UNODC Office in Peru 1999). Coca leaves are not always transformed into cocaine in the agricultural plots. Narcotics traffickers seem to take advantage of the large quantities of coca leaves transported to urban areas, ostensibly for traditional user markets. This coca is then purchased and processed at hidden facilities in urban areas near the Bolivian border. In this way the risk of being caught by authorities is reduced. From Bolivia the cocaine is dispatched to Brazil and Europe (Garcia and Antezana 2009).

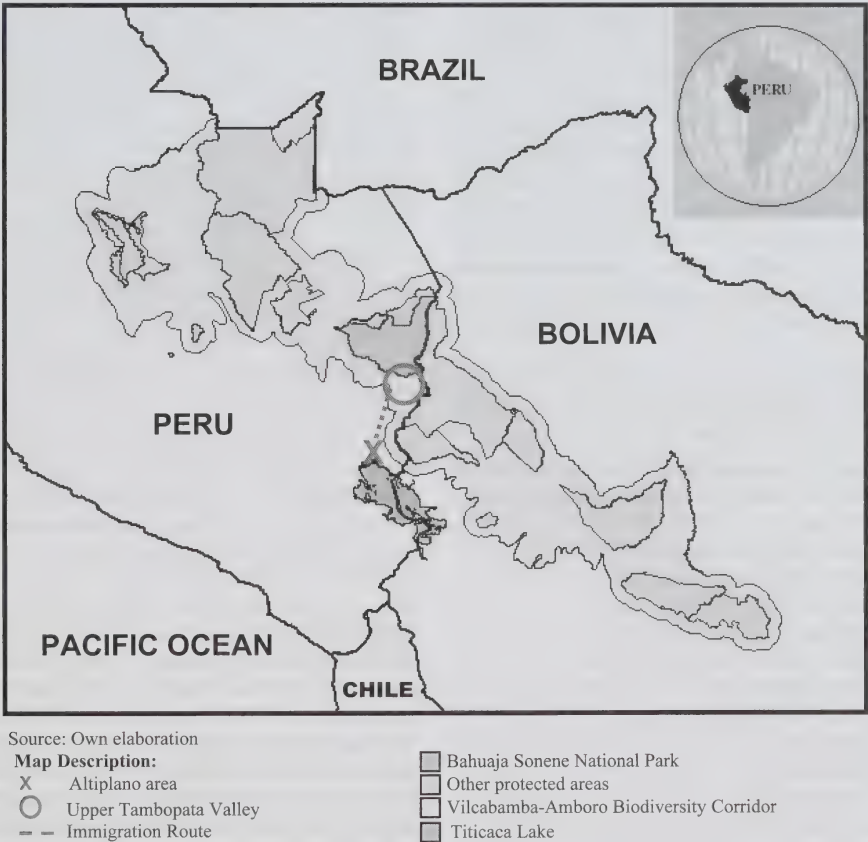


Figure 1 Map of the research area

Coca cultivation does not necessarily translate into better quality of life for the farmers in South America (Davalos, *et al.* 2008). According to the last population census, the living conditions in San Pedro de Putina Punco (SPPP), the district located in the heart of the Upper Tambopata valley, are difficult: 72% of the houses are rammed earth constructions, 88% have dirt floors, 16% have public electricity, 12% have public water, and only 9% have access to public sewage (INEI 2007). This situation is common in the major coca growing areas in Peru, where 70% of the inhabitants continue to live in poverty, and 42% in extreme poverty (Commission on Narcotic Drugs 2005).

2.2 Data collection strategies and lessons learned

A feasibility study to test if farmers would answer coca-related questions was conducted in December 2007. The pilot study for the designed questionnaire took place in May 2008, and the final survey was conducted between June and August 2008. The feasibility and pilot studies and the final survey were focused on the farmers located in San Pedro de Putina Punco (SPPP), a district in the upper Tambopata valley which is located in the deepest rainforest. All the farmers in the research area produce coffee as cash crop and some supplement their income with coca cultivation. There are five coffee co-operatives in SPPP. Farmers have to become a member of one of these co-operatives in order to be able to sell their coffee, because restrictions to coffee intermediaries are in place. The final survey was only conducted among the members of four of these co-operatives because most of the members of the remaining co-operative are based in San Juan del Oro, a district outside the research area.

The final survey consisted of a structured questionnaire which focused on agricultural production and social capital. The questionnaire was comprised of 15 sections:

1. General information about the farmer and household
2. General information about the agricultural plot and coffee area
3. Additional economic activities
4. Organic certification information
5. Cognitive social capital and identity
6. Information and communication
7. Personal aspirations and risk attitudes
8. Structural social capital
9. Covariant and idiosyncratic shocks
10. Human capital
11. Social networks
12. Coca use traditions
13. Detailed agricultural production costs
14. Labor access
15. Additional questions

The sensitive question related items of the survey are presented in the Appendix 1.

Asking farmers about their coca growing area is a sensitive question. Farmers who cultivate large areas of coca fear that the information provided could be accessed by authorities responsible for eradication programs. Thus, they might have concerns about the possible consequences of giving a truthful answer should the information become known to a third party. In these cases, the farmers need to be assured anonymity. Farmers could also be tempted to provide socially desirable answers to the interviewers. Coca has become an important focal symbol in the indigenous population's struggle for self-determination (Office of Technology Assessment 1993). Coca "yes", cocaine "no" constitutes the slogan of indigenous people (Henman 1990); the formulation tries to clearly separate traditional uses ("coca") from narcotics trafficking ("cocaine"). Hence, traditional uses such as coca chewing are ethnicity symbols (Allen 1981) and their persistence could be related to feelings of nationalism in Peru (Henman 1990). In this sense, it could be expected that farmers would not find it very problematic to indicate that they grow coca, as long as they can associate it with traditional uses. On the other hand, due to the association of larger production areas with illegal activities, coca growers may underreport the total extent of their coca production areas in an attempt to give the impression that they are growing only for traditional use.

Several strategies can help to reduce the potential biases associated with question sensitivity, item and unit nonresponse and deliberate misreporting. These strategies include: confidentiality assurances; careful selection of the data collection mode and setting of the sensitive question format; and tailoring interviewer characteristics and behavior (see Coutts and Jann 2008; Tourangeau and Yan 2007). Further information on the implementation of these strategies in our case study is provided below.

Establishing trust, and anonymity assurances

Farmers in coca growing areas tend to distrust external people. In this particular area, we found out that they trust the coffee co-operative directors. One of the directors of the coffee co-operatives signed a letter of presentation authorizing our research related to agricultural cultivation. The letter was shown to the farmers prior to conducting the survey. A pilot test conducted with and without the presentation letter demonstrated that the letter was important to reduce survey participation refusals. In the survey introduction, it was also indicated by the interviewer that the co-operative director authorized the survey because the director expected the results to benefit co-operative members. In addition, farmers were clearly told at the beginning of the survey that the data collected would remain confidential, and the academic purpose of the questionnaire was

high-lighted (see Appendix 1a). This anonymity assurance was short and precise in order to minimize suspicion among farmers as suggested by Singer, Hippler and Schwarz (1992). Coca growing was treated as a common and ordinary behavior in the research region, and a long and elaborate confidentiality assurance might have aroused farmers' reservations instead of alleviating them. A brief reminder of the assurance of confidentiality was included in the middle of the questionnaire, before the questions related to traditional coca uses and prior to the sensitive question on the coca area. The reminder stated: "In this part of the survey, we will ask questions about coca uses and cultivation. Please remember that the survey is anonymous and that there are no correct or incorrect answers" (See Appendix 1b). This follows Willis (2005) who mentions that it is important to have warm-up questions and an announcement of the switching to the sensitive topic to reduce resistance to answer.

Data collection mode

Paper and pencil self-administration as data collection method was initially considered to try to reduce interviewer bias. However, during the feasibility study, it became evident that many farmers, even those with above elementary school education (52% of the population; INEI 2007), were not able to read effortlessly. Farmers work in their fields almost all day long and do not have many opportunities to practice their reading skills. Similarly, audio computer-assisted self-interviewing (ACASI) the method of choice for collecting data on sensitive topics in developed countries (Mensch, Hewett and Erulkar 2003), was out of the scope of this project due to the lack of equipment and power supply, and the computer illiteracy in the research area. The use of computers was likely to have increased the anxiety and suspicion about the survey as described in the African situation by Mensch, *et al.* (2003). Therefore, a face-to-face interview was the data collection mode selected and emphasis was placed on the selection of interviewers, their training and behavior.

Selection of interviewers, training, and interviewers' behavior

One problem with the selection of the interviewers was the lack of sufficiently educated professionals in the research area. Thus, a group of ten students from the nearest public university, located 16 hours away from the research area, was chosen as interviewers. All of the interviewers had Aymara or Quechua ethnic backgrounds; this was an attempt to partially match interviewer-respondent characteristics. It was thought that this could increase the likelihood of participation because the matching was likely to increase trust and sympathy between the interviewer and the respondent (Tourangeau and Yan 2007). The interviewers

presented themselves as students from the local university, and no additional information was given about any university or organization outside of the country financing the study to avoid potential misunderstandings and reduce distrust among the respondents. During the pilot study, some farmers had indicated concerns about externally financed coca eradication programs and therefore references to external institutions were minimized. As a result, only partial information was given to the respondents. This is unconventional, but under the specific circumstances of the study, there was no other alternative without facing potential security problems.

For training, the interviewers first attended a two-day workshop in Puno city, followed by a three-day workshop in the research area. The same group of interviewers also conducted the pilot study to test the questions and questionnaire with the objective of identifying comprehension, recall, judgement and acceptability issues in the survey, and allowing rephrasing, eliminating or adding questions. The pilot study also allowed assessment of the performance of the interviewers, and in some cases identified areas requiring tailored training based on the feedback on performance. For example, at the beginning one of the interviewers was hesitant about asking the coca-related question and that interviewer obtained a higher than average number of nonresponses to the sensitive question. After tailored training, the interviewer was able to modify their interviewing approach.

Format of the sensitive question

The question format presupposed the sensitive behavior under study, as suggested by Tourangeau and Yan (2007). Therefore, farmers were not first asked if they had any coca areas, and then asked for the total extent of their coca areas. Instead, all farmers were directly requested to state the total extent of their coca areas ("What is your coca growing area in meters or hectares?"). However, it was found during the pilot study that the farmers did not feel comfortable with this question format and they either skipped the question or simply withdrew from the survey. As a consequence, the question format was changed and a forgiving wording was used instead. Farmers were asked: "How many 'little bushes of coca' do you have in your agricultural plot?" Thus, the farmer could answer "Only a little, I have... coca bushes". Even though a difference was hardly perceptible, with the former question it was more difficult for the farmers to start their answers with "Only a little...". So, using the latter question, it was easier for the farmers to add apologetic explanations to their answers making them feel more relaxed. This latter sensitive question format also had the advantage of employing a familiar wording for the Aymara who commonly use diminutives in their daily conversations. On the other hand, this question format might indirectly

imply that the interviewer expected that the respondent had a small number of coca bushes likely resulting in underreporting. Consequently, while nonresponses were avoided using this latter question format, underreporting was still expected to some extent.

Time period for conducting the survey and data collection setting

The farmers' agricultural plots are scattered in the mountainous Amazon rainforest in Peru. It was difficult to reach individual farmers on their agricultural plots for the survey. Therefore, to conduct the survey, we mainly took advantage of the Saint Peter's Day celebration and the General Assembly meetings of the co-operatives in June and August 2008 respectively, when the farmers congregated in the town square. Attendance to the General Assembly meetings is mandatory for all co-operative members so all of the targeted respondents would have been accessible at those events. The only way to reach or exit the town square is through an unpaved road. To take advantage of this, the survey was conducted in a large tent that was erected on the unpaved road on those key days. This tent had ten divisions, one for each pair of interviewer and respondent. Absolute privacy was not enforced because during the pilot study, it was found that farmers did not feel comfortable being the "only one" who was being interviewed; they preferred to see others doing the same. However, farmers were not able to overhear other farmers' responses. Given that all farmers have to use the same unpaved road to reach the town square regardless of their specific geographic location, potential geographical biases, which in turn can be related to important variables such as farm size and income, were likely minimized in this research.

Sampling representativeness

A convenience sampling method was applied, but at the end of the survey, we asked the farmers for their co-operative registration number and used the co-operative registration lists to infer the sample's representativeness. The co-operative registration number provided by the farmer was written on separate piece of paper and was not attached to the respondent's questionnaire. Respondents were informed about this procedure and were able to witness the procedure.

The four co-operatives under study have 3,265 members in SPPP. Table 2 shows the number of respondents per co-operative. The number of collected questionnaires amounted to 508. In total, 12 respondents were excluded from the sample because their co-operative registration number was missing. In two cases, the farmers had refused to provide this information and in ten cases, the interviewers had forgotten to ask the respondents about their registration

number at the end of the interview. Therefore the absence of information was more associated with interviewer error than with the farmers' unwillingness to provide this information.

Table 2
Number of respondents per co-operative

	Total Number of Co-operative Members in SPPP	Survey's Sample Size	Percentage of Co-operative Members Interviewed (%)
Co-operative 1	756	106	14
Co-operative 2	911	138	15
Co-operative 3	887	138	16
Co-operative 4	711	114	16
Total	3,265	496	15

Source: Own survey.

In order to test for representativeness of the sample, the distribution of the co-operative registration numbers obtained from the survey sample was compared with the distribution of the co-operative registration numbers from a simulated simple random sample without replacement obtained from co-operative lists. The co-operative lists were ordered by the registration number of the co-operative members and co-operative registration numbers are associated with the members' date of registration. Thus, most of the older farmers have lower registration numbers and the younger farmers have higher ones. Unfortunately, the co-operatives did not have other membership data available such as total land, coffee or coca hectares that might be used to select a stratified random sample. Two types of tests were used for comparison of the samples: a two-sample Wilcoxon rank-sum (Mann-Whitney) test and a two-sample Kolmogorov-Smirnov test for equality of distribution functions. The first test assesses how probable it is that the two groups come from the same distribution, and assumes that differences observed are caused by chance fluctuation. The second test is similar to the first one, but in addition it is sensitive to differences in both the location and shape of the empirical cumulative distribution functions of the two groups. The results of both tests failed to reject the null hypothesis of equality of distribution between the survey sample and the simulated simple random sample at a significance level of 0.05. Thus, the results suggest that the survey sample is equivalent to a simple random sample, and therefore representative of the population under study.

3. Survey results and validation issues

3.1 Survey results

The survey response rate was around 90%, which is well above the minimum recommended response rate of 60% (Punch 2003). From the 496 completed questionnaires, 19

respondents (less than 4%) did not answer the coca-related question. When comparing the descriptive statistics of socio-economic, institutional, and coca-related variables, there were some significant differences between all the observations (without the non-respondents) and the 'sensitive question non-respondents' (see Appendix 2). The sensitive question non-respondents were all male, with a larger percentage of Aymara ethnic background, and more children. In addition, a larger percentage of them used coca as medicine. Interestingly, significantly more non-respondents are highly risk averse (73.7%) compared to all the other respondents (28.6%). This could indicate a potential fear of the 'sensitive question non-respondents' of interviewer disclosure of information to third parties. The setup of the risk aversion test followed by Binswanger (1980) is presented in Appendix 1c.

Basic comparative descriptive statistics of coca and non coca growers are presented in Table 3. The number of valid questionnaires was 477, if we do not account for the non respondents of the sensitive question. Of them, 64% indicated that they are coca growers.

There are no statistically significant differences with respect to general socio-economic characteristics (age, sex,

ethnic group, and number of children) between coca and non-coca growers. The only difference was observed in education. Non-coca growers have more years of schooling than coca growers. Coca growers have less total and primary forest areas, and more fallow land than non coca growers, although these differences are not statistically significant. Coca and non-coca growers have similar coffee and staple food areas. On the contrary, coca growers and non-coca growers show statistically significant differences in the social capital variables. More non-coca growers than coca growers find it important to obey national law. On the other hand, less non-coca growers than coca growers have experienced a negative change in trust towards their neighbors during the last five years, and have worked in community activities during the last year.

There is a statistically significant relationship between coca growing and traditional uses. A higher percentage of coca growers than non-coca growers chew coca and uses coca as medicine. More importantly, more coca growers find it easier to sell coca leaves than non-coca growers in the hypothetical case that they would cultivate coca for commercial purposes.

Table 3
Comparative descriptive statistics between coca and non coca growers

Variable	Coca Growers	Non Coca Growers
Age	42.5 (12.7)	41.7 (12.5)
Male (%)	93.9	94.9
Aymara (%)	81.4	82.5
Number of Children	3.0 (2.0)	2.9 (2.1)
Years of schooling	8.2* (3.3)	8.7* (3.3)
Total area (ha)	7.9 (8.4)	8.0 (7.8)
Coffee area (ha)	2.2 (2.0)	2.2 (1.4)
Area secondary forest (fallow area)	1.6 (2.4)	1.4 (2.1)
Primary forest area (ha)	3.9 (7.5)	4.2 (7.0)
Staple food area (ha)	0.5 (0.7)	0.5 (0.6)
No other economic activities (%)	46.8	48.9
High risk aversion (%)	30.5	25.3
Important to obey national laws (%)	81.9**	88.6**
Negative change in trust in the last 5 years (%)	19.3**	12.5**
Have worked in community activities in 2007 (%)	92.0**	84.7**
Farmer chews coca (%)	76.0***	53.1***
Farmer uses coca as medicine (%)	81.7***	54.8***
Perception that it is easy to sell coca leaves (%)	26.4**	18.5**
Number of coca bushes	3,093 (6,710)	-
Number of Observations	305	172

Standard deviations are in parentheses for continuous variables.

Coca Growers and Non Coca Growers means are statistically different (T-test with unequal variances) at:

* 0.1 significance level, ** 0.05 significance level, *** 0.01 significance level.

Source: Own calculations.

Finally, it is important to mention that the average number of coca bushes is relatively low, which could be due to underreporting of commercial coca growing areas or to coca cultivation only for self-consumption, or both. It is not possible to distinguish between those two scenarios, which makes it easier for commercial coca growers to disguise themselves as coca growers who produce for traditional uses.

3.2 Validation issues

The validity of individual responses cannot be verified directly because there is little prior empirical research on this topic, and there is an absence of other sources of confirming data. However, it is possible to provide a rough comparison between the survey data and the total area of coca production recounted by international organizations for the upper Tambopata valley using satellite data. The United Nations Office on Drugs and Crime (UNODC 2009) indicates that 940 hectares of coca were cultivated in the upper Tambopata valley in 2008. The conventional coca cultivation density for regions with traditional coca growers could be between 35,000 and 40,000 bushes per hectare (UNODC 2001) (During the 90s, the coca cultivation density was lower, between 20,000 and 25,000 bushes per hectare (UNODC 2009)). The coca cultivation density in the particular valley is relatively low because coca growers intercrop coca with coffee and staples, although the yields per bush have increased during the last years (UNODC 2009). Therefore, it is expected that the total number of coca bushes for this valley would be approximately from 32.9 to 37.6 million.

Our sample of 477 respondents (excluding farmers who did not report their co-operative registration number and non respondents to the sensitive question) reported a total of 960,000 coca bushes. This sample corresponds to 14.6% of a total of 3,265 co-operative members in SPPP. Thus, extrapolating for the total number of co-operative members located in the SPPP district would result in a total of 6.6 million coca bushes. In addition, we need to consider that the upper Tambopata valley also includes San Juan del Oro district which has around the same population as SPPP district (INEI 2007). Under the very strong assumption that farmers in SPPP behave similarly to the farmers in San Juan del Oro - at least in terms of coca cultivation - this would double the number of coca bushes for the entire upper Tambopata valley to around 13.2 million. This last estimate is between 35 and 40% of the 32.9 to 37.6 million obtained from UNODC satellite data. This result is in the expected range of reporting on sensitive issues. For reporting on abortion, this range is between 35 to 59% (Fu, Darroch, Henshaw and Kolb 1998), and for the use of opiates or cocaine between 30 to 70% (Tourangeau and Yan 2007).

4. Summary and conclusions

Coca, a raw material for the production of cocaine, is cultivated in Colombia, Peru and Bolivia. In the latter two countries, traditional uses of coca by indigenous populations date back to around 3000 B.C. (Rivera, *et al.* 2005). Nevertheless, asking farmers about the extent of their coca cultivation areas is considered a sensitive question. Coca growers are afraid of eradication programs even if they do not sell coca to the narcotics traffic business because it is difficult to distinguish between coca growers whose production is commercially oriented and those who produce only for self-consumption. Thus, farmers tend not to participate in surveys, not to answer any sensitive questions, or to underreport their coca cultivation areas in an attempt to minimize their identification for possible eradication.

Against this background, household-level data collection procedures need to consider and evaluate strategies to reduce nonresponses and misreporting. Most of the strategies used in our research area in Peru were based on best practices reported in the literature review. Some of the strategies that worked in our case were establishment of trust with the farmers using a presentation letter from a coffee co-operative director, confidentiality assurance at the beginning and in the middle of the questionnaire, matching of interviewer-respondent ethnic background characteristics, training of interviewers to reduce their hesitance to ask sensitive questions, changing the format of the sensitive question to a familiar and forgiving wording, and non enforcement of absolute privacy to prevent each farmer from feeling that they were the "only one" who was interviewed.

The validity of farmers' individual responses on their coca area extensions cannot be checked because the topic has produced little prior empirical research, and there is an absence of other sources of household-level confirming data. Thus, the extent of misreporting was evaluated using aggregate data. The results suggest that farmers only reported between 35 to 40% of their actual coca areas. Still, those values are between the ranges of what could be expected for answers to sensitive questions. In terms of survey nonresponse and sensitive question nonresponses, the results were more encouraging indicating values of 10% and of around 4%, respectively.

When conducting the survey, we mainly took advantage of celebrations and co-operative General Assemblies for which farmers congregated in town, since farmers are otherwise highly dispersed in the rainforest. The survey followed a convenience sampling method but it was possible to test the representativeness of this sample because all of the farmers are registered in one of the co-operatives in the research area. The obtained sample was compared with a simulated simple random sample without replacement

where each farmer had the same probability to be selected by chance from the co-operative member lists. There were no statistical differences in the distribution functions, so the sample is equivalent to a simple random one. The main drawback of this approach is that after the interview, we needed to ask the respondents for their co-operative member number. Even though the respondents were told that the co-operative identification number was not attached to their questionnaires, some farmers might have had doubts about it, and this could have had effects on confidentiality assurance credibility in following interviews due to word spreading.

On the other hand, comparing the characteristics of non-respondents to sensitive questions with the rest of respondents indicates that non-respondents were highly risk averse. Even though the number of non-respondents was small (less than 4% of the total sample), this could suggest that the main reason for item non-reporting is the fear of the consequences of the information leaking to third parties.

The coca areas reported by the farmers were on average very small. This could be an attempt by commercial coca growers to appear to be cultivating only for self-consumption. Coca growing for traditional uses does not have a negative connotation *per se* given that it is a symbol of ethnicity and the indigenous population's struggle for self-determination (Office of Technology Assessment 1993). It is not possible to distinguish farmers who underreported the extent of their coca cultivation areas from those who grow coca for self-consumption. Unfortunately, commercial coca growers can take advantage of this situation to continue growing coca under the guise of traditional uses.

Acknowledgements

The research was funded by BMZ (the Federal Ministry for Economic Cooperation and Development, Germany) through the DAAD (German Academic Exchange Service), and by LACEEP (Latin American and Caribbean Environmental Economics Program).

Appendix 1

Relevant parts of the questionnaire

A) Presentation:

Good morning/afternoon/night. My name is _____. I am a student at _____. We are conducting a survey to identify the risks and vulnerabilities of coffee producers in your community. The coffee co-operative directives are aware of this survey and believe that the result could benefit the community. If you decide to answer our questionnaire, you may skip any questions or withdraw from this study at any time. The data collected in this survey will remain CONFIDENTIAL and will be used only for ACADEMIC purposes. Your answers and opinions are extremely important for the co-operative and us. Would you be prepared to respond to some questions?

a) Yes (proceed)

b) No (thank the respondent, withdraw the survey, and indicate the characteristics of the person in format 1)

B) Coca Related Questions:

In this part, we will ask about coca uses and cultivation. Please, remember that this survey is anonymous and that there are no correct or incorrect answers.

Do you chew coca leaves?

a) Yes b) No

Do you use coca leaves as medicine?

a) Yes b) No

Do you feel obligated to offer coca leaves to your guests during ayni and minka activities?

a) Yes b) No

Do you use coca leaves for rituals?

a) Yes b) No

Do you use coca leaves for payment to external workers?

a) Yes b) No

Do you use coca leaves as product exchange or as a gift for friends and relatives?

a) Yes b) No

How many little bushes of coca do you have in your agricultural plot?

C) Risk Aversion Question:

This is a game. Before playing it, you need to choose one of the options displayed below. Then I toss a coin. If for example you have chosen option H, and I toss the coin and it is heads, you do not win any money at all; but if it is tails, you win S/.200. On the other hand, if you have chosen option A, you receive S/.50 regardless of if the tossed coin is heads or tails. Which option from all of the above would you choose before I toss the coin?

OPTION	If it is heads, you win:	If it is tails, you win:
A	50 soles	50 soles
B	45 soles	95 soles
C	40 soles	120 soles
D	35 soles	125 soles
E	30 soles	150 soles
F	20 soles	160 soles
G	10 soles	190 soles
H	0 soles	200 soles

Appendix 2

Comparative descriptive statistics between all observations and sensitive question non respondents

Variable	All Observations ^a	Sensitive Question Non Respondent
Age	42.2 (12.6)	45.9 (9.9)
Male (%)	94.3***	100***
Aymara (%)	81.8**	94.7**
Number of Children	3.0** (2.0)	4.1** (2.0)
Years of schooling	8.4 (3.3)	7.5 (2.9)
Total area (ha)	7.9 (8.3)	6.8 (3.2)
Coffee area (ha)	2.2 (1.8)	2.5 (1.2)
Area secondary forest (fallow area)	1.6 (2.3)	1.4 (1.1)
Primary forest area (ha)	4.0 (7.3)	2.9 (3.3)
Staple food area (ha)	0.5 (0.7)	0.6 (0.6)
No other economic activities (%)	47.5	57.9
High risk aversion (%)	28.6***	73.7***
Important to obey national laws (%)	84.3	89.5
Negative change in trust in the last 5 years (%)	16.8	26.3
Have worked in community activities in 2007 (%)	89.4	89.5
Farmer chews coca (%)	67.7	73.7
Farmer uses coca as medicine (%)	72.0*	84.2*
Easy to sell coca leaves (%)	23.6	27.8
Number of Observations	477	19

Standard deviations are in parentheses for continuous variables.

a) All observations without sensitive question non respondents.

Non respondent means are statistically different from the entire sample (T-test with unequal variances) at:

* 0.1 significance level, ** 0.05 significance level, *** 0.01 significance level.

Source: Own calculations.

References

- Allen, C. (1981). To be Quechua: The symbolism of coca chewing in highland Peru. *American Ethnologist*, 8, 1, 157-171.
- Barnett, J. (1998). Sensitive questions and response effects: An evaluation. *Journal of Managerial Psychology*, 13, 1/2, 63-67.
- Bedoya, E. (2003). Estrategias productivas y el riesgo entre los cocaleros del valle de los ríos apurímac y ene. In *Amazonia: Procesos Demográficos y Ambientales*, (Eds., C. Aramburu and E. Bedoya), Consorcio de Investigación Económica y Social. Lima, Peru.
- Binswanger, H. (1980). Attitude towards risk: Experimental measurement in rural India. *American Journal of Economics*, 62, 395-407.
- Botvin, G., Griffin, K., Diaz, T., Scheier, L., Williams, C. and Epstein, J. (2000). Preventing illicit drug use in adolescents: Long-term follow-up data from a randomized control trial of a school population. *Addictive Behaviors*, 25, 5, 769-774.
- Caballero, V., Dietz, E., Taboada, C. and Anduaga, J. (1998). Diagnostico Rural Participativo de las Cuencas Alto Inambari y Alto Tambopata Provincia de Sandia, Departamento de Puno. GTZ. Lima, Peru.
- Caulkins, J., Reuter, P., Iguchi, M. and Chiesa, J. (2005). How goes the War on Drugs? An Assessment of U.S. Drug Problems and Policy. RAND Drug Policy Research Center. U.S.
- Collins, J. (1984). The maintenance of peasant coffee production in a peruvian valley. *American Ethnologist*, 11, 3, 413-438.
- Commission on Narcotic Drugs (2005). Alternative Development: A Global Thematic Evaluation. Final Synthesis Report. Forty - Eight Session E/CN.7/2005/CRP.3. Austria.
- Coutts, E., and Jann, B. (2008). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). ETH Zurich Sociology, Working Paper, No. 3.
- Davalos, L., Bejarano, A. and Correa, L. (2008). Disabusing cocaine: Pervasive myths and enduring realities of a globalised commodity. *International Journal of Drug Policy*, 20, 5, 381-386.

- Davis, C., Thake, J. and Vilhena, N. (2009). Social Desirability Biases in Self-Reported Alcohol Consumption and Harms. Addictive Behaviors. Article in Press.
- Durand, F. (2005). El Problema Cocalero y el Comercio Informal para Uso Tradicional. Debate Agrario 39. Lima, Peru.
- Fergusson, D., Boden, J. and Horwood, L. (2008). The developmental antecedents of illicit drug use: Evidence from a 25-Year longitudinal study. *Drug and Alcohol Dependence*, 96, 165-177.
- Fu, H., Darroch, J., Henshaw, S. and Kolb, E. (1998). Measuring the extent of abortion underreporting in the 1995 National Survey of Family Growth. *Family Planning Perspectives*, 30, 3, 128-138.
- García, J., and Antezana, J. (2009). Diagnostico de la Situación del Desvío de IQ al Narcotráfico. ConsultAndes and DEVIDA. Lima, Peru.
- Gibson, B., and Godoy, R. (1993). Alternatives to coca production in Bolivia: A computable general equilibrium approach. *World Development*, 21, 6, 1007-1021.
- Henman, A. (1990). Tradicion y represion: Dos experiencias en america del sur. In *Coca, Cocaína y Narcotráfico. Laberinto en los Andes*, (Eds., García – D. Sayan), Comision Andina de Juristas. Lima, Peru.
- Holmstedt, B., Jaatmaa, E., Leander, K. and Plowman, T. (1977). Determination of cocaine in some South American species of erythroxylum using mass fragmentography. *Phytochemistry*, 16, 1753-1755.
- INCB (2009). Report on the International Narcotics Control Board for 2009. United Nations Publication. New York, U.S.A.
- Ibanez, M., and Carlsson, F. (2010). A survey-based choice experiment on coca cultivation. *Journal of Development Economics*, 93, 2, 249-263.
- INEI (2007). Censos Nacionales 2007: XI de Población y VI de Vivienda. Lima, Peru.
- Johnson-Hanks, J. (2002). The lesser shame: Abortion among educated women in southern Cameroon. *Social Science & Medicine*, 55, 8, 1337-1349.
- Mansfield, D. (2006). Development in Drug Environment: A Strategic Approach to Alternative Development. Discussion Paper. Development Oriented Drug Control Program. GTZ. Germany.
- Mensch, B., Hewett, P. and Erulkar, A. (2003). The reporting of sensitive behavior by adolescents: A methodological experiment in Kenya. *Demography*, 40, 2, 247-268.
- Obando, E. (2006). U.S. Policy toward Peru: At odds for twenty years. In *Addicted to Failure. U.S. Security Policy in Latin America and the Andean Region*, (Eds., B. Loveman). Rowman & Littlefield Publishers Inc. U.S.
- Office of Technology Assessment (1993). Alternative Coca Reduction Strategies in the Andean Region. U.S. Congress. OTA-F-556. Washington, U.S.
- Punch, K. (2003). Survey research. The basics. *Sage Publications, Inc.* U.K.
- Rivera, M., Aufderheide, A., Cartmell, L., Torres, C. and Langsjoen, O. (2005). Antiquity of coca – Leaf chewing in the south central Andes: A 3000 year archaeological record of coca - Leaf chewing from Northern Chile. *Journal of Psychoactive Drugs*, 37, 4, 455-458.
- Rospigliosi, F. (2004). Analisis de la Encuesta DEVIDA-INEI. In *El Consumo Tradicional de la Hoja de Coca en el Peru*, (Ed., F. Rospigliosi). Instituto de Estudios Peruanos. Lima, Peru.
- Singer, E., Hippler, H. and Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research*, 4, 3.
- Strunin, L. (2001). Assessing alcohol consumption: developments from qualitative research methods. *Social Science & Medicine*, 53, 2, 215-226.
- Thoumi, F. (2003). Illegal Drugs, Economy, and Society in the Andes. Woodrow Wilson Center Press. Washington, U.S.
- Torrico, J., Pohlman, H. and Janssens, M. (2005). Alternatives for the transformation of drug production areas in the chapare region, Bolivia. *Journal of Food, Agriculture and Development*, 3, 3-4, 167-172.
- Tourangeau, R., and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 5, 859-883.
- UNODC (2001). Alternative Development in the Andean Area. The UNDCP Experience. Revised Edition. ODCCP Studies on Drugs and Crime. New York, U.S.
- UNODC (2009). Perú. Monitoreo de Cultivos de Coca 2008. Lima, Peru.
- UNODC (2011). Perú. Monitoreo de Cultivos de Coca 2010. Lima, Peru.
- UNODC Office in Peru (1999). Desarrollo Alternativo del Inambari y Tambopata. Documento de Proyecto AD/PER/99/D96. Availability: <http://www.onudd.org.pe/web/Html/Templates/proyectos.htm> (accessed on June 15, 2009).
- U.S. Department of State (2009). International Narcotics Control Strategy Report. Volume I: Drug and Chemical Control. Bureau for International Narcotics and Law Enforcement Affairs. U.S.
- Varkey, P., Balakrishna, P., Prasad, J., Abraham, S. and Joseph, A. (2000). The reality of unsafe abortion in a rural community in South India. *Reproductive Health Matters*, 8, 16, 83-91.
- Willis, G. (2005). Cognitive interviewing. A tool for improving questionnaire design. *Sage Publications, Inc.* U.S.
- Zufferey, A., Michaud, P., Jeannin, A., Berchtold, A., Chossis, I., van Melle, G. and Suris, J. (2007). Cumulative risk factors for adolescent alcohol misuse and its perceived consequences among 16 to 20 year old adolescents in Switzerland. *Preventive Medicine*, 45, 2-3, 233-239.

Imputation for nonmonotone nonresponse in the survey of industrial research and development

Jun Shao, Martin Klein and Jing Xu¹

Abstract

Nonresponse in longitudinal studies often occurs in a nonmonotone pattern. In the Survey of Industrial Research and Development (SIRD), it is reasonable to assume that the nonresponse mechanism is past-value-dependent in the sense that the response propensity of a study variable at time point t depends on response status and observed or missing values of the same variable at time points prior to t . Since this nonresponse is nonignorable, the parametric likelihood approach is sensitive to the specification of parametric models on both the joint distribution of variables at different time points and the nonresponse mechanism. The nonmonotone nonresponse also limits the application of inverse propensity weighting methods. By discarding all observed data from a subject after its first missing value, one can create a dataset with a monotone ignorable nonresponse and then apply established methods for ignorable nonresponse. However, discarding observed data is not desirable and it may result in inefficient estimators when many observed data are discarded. We propose to impute nonrespondents through regression under imputation models carefully created under the past-value-dependent nonresponse mechanism. This method does not require any parametric model on the joint distribution of the variables across time points or the nonresponse mechanism. Performance of the estimated means based on the proposed imputation method is investigated through some simulation studies and empirical analysis of the SIRD data.

Key Words: Bootstrap; Imputation model; Kernel regression; Missing not at random; Longitudinal study; Past-value-dependent.

1. Introduction

Longitudinal studies, in which data are collected from every sampled subject at multiple time points, are very common in research areas such as medicine, population health, economics, social sciences, and sample surveys. The statistical analysis in a sample survey typically aims to estimate or make inference on the mean of a study variable at each time point. Nonresponse or missing data in the study variable is a serious impediment to performing a valid statistical analysis, because the response propensity (PSI) may directly or indirectly depend on the value of the study variable. Nonresponse is monotone if, whenever a value is missing at a time point t , all future values at $s > t$ are missing. We focus on nonmonotone nonresponse, which often occurs in longitudinal surveys. In the Survey of Industrial Research and Development (SIRD) conducted jointly by the U.S. Census Bureau and the U.S. National Science Foundation (NSF), for example, a business may be a nonrespondent on research and development expenditures at year $t - 1$ but a respondent at year t . For ease we refer to SIRD in the present tense throughout, but we note that as of 2008, it has been replaced by the Business R&D and Innovation Survey.

Some existing methods for handling nonmonotone nonresponse can be briefly described as follows. The parametric approach assumes parametric models for both the PSI and

the joint distribution of the study variable across time points (e.g., Troxel, Harrington and Lipsitz 1998, Troxel, Lipsitz and Harrington 1998). The validity of the parametric approach, however, depends on whether parametric models are correctly specified. Vansteelandt, Rotnitzky and Robins (2007) proposed some methods under some models of the PSI at time t conditional on observed past data. Xu, Shao, Palta and Wang (2008) derived an imputation procedure under the assumptions that (i) the PSI at t depends only on values of the study variable at time $t - 1$ and (ii) the study variables over different time points is a Markov chain. Another approach, which will be referred to as censoring, is to create a dataset with “monotone nonresponse” by discarding all observed values of the study variable from a sampled subject after its first missing value. Methods appropriate for monotone nonresponse (e.g., Diggle and Kenward 1994, Robins and Rotnitzky 1995, Paik 1997) can then be applied to the reduced dataset. This approach may be inefficient when many observed data are discarded. Furthermore, in practical applications it is not desirable to throw away observed data.

The purpose of this article is to propose an imputation method for longitudinal data with nonmonotone nonresponse under the past-value-dependent PSI assumption described by Little (1995): at a time point t , the nonresponse propensity depends on values of the study variable at time points prior to t . This assumption on the PSI is weaker than

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706. E-mail: shao@stat.wisc.edu; Martin Klein, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C. 20233; Jing Xu, Department of Statistics, University of Wisconsin, Madison, WI 53706.

that in Xu *et al.* (2008) and is different from those in Vansteelandt *et al.* (2007). We consider imputation which does not require building a model for the PSI. Imputation is commonly used to compensate for missing values in survey problems (Kalton and Kasprzyk 1986). Once all missing values are imputed, estimates of parameters are computed using the estimated means for complete data by treating imputed values as observations. The proposed imputation and estimation methodology, including a bootstrap method for variance estimation, is introduced in Section 2. To examine the finite sample performance of the proposed method, we present some simulation results in Section 3. We also include an application of the proposed method to the SIRD. The last section contains some concluding remarks.

2. Methodology

We consider the model-assisted approach for survey data sampled from a finite population P . We assume that the population P is divided into a fixed number of imputation classes, which are typically unions of some strata. Within each imputation class, the study variable from a population unit follows a superpopulation. Let y_t be the study variable at time point t , $t = 1, \dots, T$, $\mathbf{y} = (y_1, \dots, y_T)$, δ_t be the indicator of whether y_t is observed, and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_T)$. Since imputation is carried out independently within each imputation class, for simplicity of notation we assume in this section that there is only a single imputation class.

Throughout this paper, we consider nonmonotone nonresponse and assume that there is no nonresponse at baseline $t = 1$. The PSI is past-value-dependent if

$$P(\delta_t = 1 \mid \mathbf{y}, \delta_1, \dots, \delta_{t-1}, \delta_{t+1}, \dots, \delta_T) \\ = P(\delta_t = 1 \mid y_1, \dots, y_{t-1}, \delta_1, \dots, \delta_{t-1}), \quad t = 2, \dots, T, \quad (1)$$

where P is with respect to the superpopulation. When nonresponse is monotone, the past-value-dependent PSI becomes ignorable (Little and Rubin 2002), since we either observe all past values or know with certainty that y_t is missing if it is missing at $t - 1$, and an imputation method using linear regression proposed by Paik (1997) can be used. When nonresponse is nonmonotone, however, the past-value-dependent PSI is nonignorable because the response indicator at time t is statistically dependent upon previous values of the study variable, some of which may not be observed. In this case Paik's method does not apply.

2.1 Imputation for subjects whose first missing is at t

Let $t > 1$ be a fixed time point and $r + 1$ be the time point at which the first missing value of \mathbf{y} occurs. When $r + 1 = t$, i.e., a subject whose first missing value is at t ,

our proposed imputation procedure is the same as that for the case of monotone nonresponse (Paik 1997). However, we still need to provide a justification since we have a different PSI. It is shown in the Appendix that, under assumption (1),

$$E(y_t \mid y_1, \dots, y_{t-1}, \delta_1 = \dots = \delta_{t-1} = 1, \delta_t = 0) \\ = E(y_t \mid y_1, \dots, y_{t-1}, \delta_1 = \dots = \delta_{t-1} = 1, \delta_t = 1) \quad t = 2, \dots, T, \quad (2)$$

where E is the expectation with respect to the superpopulation. Denote the quantity on the first line of (2) by $\phi_{t,t-1}(y_1, \dots, y_{t-1})$, which is the conditional expectation of a missing y_t given observed y_1, \dots, y_{t-1} . If $\phi_{t,t-1}$ is known, then a natural imputed value for y_t is $\phi_{t,t-1}(y_1, \dots, y_{t-1})$. However, $\phi_{t,t-1}$ is usually unknown. Since $\phi_{t,t-1}$ cannot be estimated by regressing y_t on y_1, \dots, y_{t-1} based on data from subjects with missing y_t values, we need to use (2), i.e., the fact that $\phi_{t,t-1}$ is the same as the quantity on the second line of (2), which is the conditional expectation of an observed y_t given observed y_1, \dots, y_{t-1} and can be estimated by regressing y_t on y_1, \dots, y_{t-1} , using data from all subjects having observed y_t and observed y_1, \dots, y_{t-1} . Note that (2) is a counterpart of (5) in Xu *et al.* (2008) under the last-value-dependent assumption, which is stronger than the past-value-dependent assumption (1). Under a stronger assumption, we are able to utilize more data in regression fitting.

Suppose that a sample S is selected from P according to a given probability sampling plan. For each $i \in S$, $\delta_i = (\delta_{i1}, \dots, \delta_{iT})$ is observed, the study variable y_{it} with $\delta_{it} = 1$ is observed, and y_{it} with $\delta_{it} = 0$ is not observed, $t = 1, \dots, T$. With respect to the superpopulation, $(\mathbf{y}_i, \boldsymbol{\delta}_i)$ has the same distribution as $(\mathbf{y}, \boldsymbol{\delta})$ and $(\mathbf{y}_i, \boldsymbol{\delta}_i)$'s are independent, where $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$. For $t = 2, \dots, T$, let $\hat{\phi}_{t,t-1}$ be the regression estimator of $\phi_{t,t-1}$ based on observations with $\delta_{i1} = \dots = \delta_{i(t-1)} = 1$. A missing y_{it} with observed $y_{i1}, \dots, y_{i(t-1)}$ is then imputed by $\hat{y}_{it} = \hat{\phi}_{t,t-1}(y_{i1}, \dots, y_{i(t-1)})$.

To illustrate, we consider the case of $t = 3$ or 4. The horizontal direction in Table 1 corresponds to time points and the vertical direction corresponds to different missing patterns, where each pattern is represented by a vector of 0's and 1's with 0 indicating a missing value and 1 indicating an observed value. For $t = 3$ and $r = 2$, as the first of the two steps, we consider missing data at time 3 with first missing at time 3, i.e., pattern (1,1,0). According to imputation model (2), we fit a regression using data in pattern (1,1,1) indicated by + (used as predictors) and × (used as responses). Then, imputed values (indicated by ○) are obtained from the fitted regression using data indicated by * as predictors. For $t = 4$ and $r = 3$, imputation in pattern (1,1,1,0) can be similarly done using data in pattern (1,1,1,1) for regression fitting.

Table 1
Illustration of imputation process

Pattern	Step 1: $r = 2, t = 3$			Step 2: $r = 1, t = 3$		
	Time			Time		
	1	2	3	1	2	3
(1,0,0)				*		○
(1,1,0)	*	*	○	+		⊗
(1,1,1)	+	+	×			
(1,0,1)						

Pattern	Step 1: $r = 3, t = 4$				Step 2: $r = 2, t = 4$				Step 3: $r = 1, t = 4$			
	Time				Time				Time			
	1	2	3	4	1	2	3	4	1	2	3	4
(1,0,0,0)									*			○
(1,1,0,0)					*	*		○	+			⊗
(1,1,1,0)	*	*	*	○	+	+		⊗	+			⊗
(1,0,1,0)									*			○
(1,0,0,1)												
(1,1,0,1)												
(1,0,1,1)												
(1,1,1,1)	+	+	+	×								

+: observed data used in regression fitting as predictors.

×: observed data used in regression fitting as responses.

⊗: imputed data used in regression fitting as responses.

*: observed data used as predictors in imputation.

○: imputed values.

What type of regression we can fit to obtain \hat{y}_t ? It is shown in the Appendix that, if (1) holds and $E(y_t | y_1, \dots, y_{t-1})$ is linear in y_1, \dots, y_{t-1} for any t in the case of no nonresponse, then

$$E(y_t | y_1, \dots, y_{t-1}, \delta_1 = \dots = \delta_{t-1} = 1) \text{ is linear in } y_1, \dots, y_{t-1} \quad (3)$$

and, hence, linear regression under the model-assisted approach can be used to estimate $\phi_{t,t-1}$. If $E(y_t | y_1, \dots, y_{t-1})$ is not linear, one of the methods described in Section 2.3 can be applied.

2.2 Imputation for subjects whose first missing is at $r + 1 < t$

Imputation for a subject whose first missing value is at time $r + 1 < t$ is more complicated and very different from that for the case of monotone nonresponse. This is because when $r + 1 < t$ and nonresponse is monotone,

$$\begin{aligned} E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_t = 0) \\ = E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_t = 1) \\ r = 1, \dots, t-2, \quad t = 2, \dots, T, \quad (4) \end{aligned}$$

whereas (4) does not hold when nonresponse is non-monotone (see the proof in the Appendix). Hence, we need to construct different models for subjects whose first missing value is at $r + 1 < t$. It is shown in the Appendix that, when $r + 1 < t$,

$$\begin{aligned} E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) \\ = E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) \\ r = 1, \dots, t-2, \quad t = 2, \dots, T. \quad (5) \end{aligned}$$

We now explain how to use (5) to impute missing values at a fixed time point t . Let $\phi_{t,r}(y_1, \dots, y_r)$ be the quantity on the first line of (5). If $\phi_{t,r}$ is known, then y_t can be imputed by $\phi_{t,r}(y_1, \dots, y_r)$. Otherwise, it needs to be estimated based on (5). Unlike in model (2) or (4), the conditional expectation on the second line of (5) is conditional on a missing y_t ($\delta_t = 0$), although y_1, \dots, y_r are observed. If we carry out imputation sequentially according to $r = t-1, t-2, \dots, 1$, then, for a given $r < t-1$, the missing y_t values from subjects whose first missing is at time point $r+2$ have already been imputed using the method in this section or Section 2.1. We can fit a regression between imputed y_t and observed y_1, \dots, y_r using data from all subjects having already imputed y_t (used as responses), observed y_1, \dots, y_r (used as predictors), and $\delta_{r+1} = 1$. Once an estimator $\hat{\phi}_{t,r}$ is obtained, a missing y_{it} with first missing at $r+1$ is then imputed by $\hat{y}_{it} = \hat{\phi}_{t,r}(y_{i1}, \dots, y_{ir})$.

Consider again the case of $t = 3$ or 4 and Table 1. Following the first step for $t = 3$ discussed in Section 2.1, at the second step, we impute missing values with $r = 1$ in pattern (1,0,0). According to imputation model (5), we fit a regression using data in pattern (1,1,0) indicated by + (used as predictors) and ⊗ (previously imputed values used as

responses). Then, imputed values (indicated by \circ) are obtained from the fitted regression using data indicated by $*$ as predictors. For $t = 4$, following the first step discussed in Section 2.1, at the second step ($r = 2$) we fit a regression using data in pattern (1,1,1,0) indicated by $+$ (used as predictors) and \otimes (previously imputed values used as responses). Then, imputed values (indicated by \circ) at $t = 4$ in pattern (1,1,0,0) are obtained from the fitted regression using data indicated by $*$ as predictors. At step 3 for $t = 4$, we fit a regression using data in patterns (1,1,0,0) and (1,1,1,0) indicated by $+$ (used as predictors) and \otimes (previously imputed values used as responses). Then, imputed values (indicated by \circ) at $t = 4$ in patterns (1,0,0,0) and (1,0,1,0) are obtained from the fitted regression using data indicated by $*$ as predictors.

Although at time t , imputation has to be carried out sequentially as $r = t - 1, \dots, 1$, imputation for different time points can be done in any order. This can be seen from the illustration given by Table 1, where the imputed values at $t = 3$ are not involved in the imputation process at $t = 4$ or vice versa, although some observed data will be repeatedly used in regression fitting. When data come according to time, it is natural to impute nonrespondents in the order $t = 2, \dots, T$.

Why can we use previously imputed values as responses in the estimation of the regression function $\phi_{t,r}$ when $r < t - 1$? For given t and $r < t - 1$, a previously imputed value with first missing at $s + 1 > r + 1$ is an estimator of

$$\begin{aligned}\tilde{y}_t &= E(y_t | y_1, \dots, y_s, \delta_1 = \dots = \delta_s = 1, \delta_{s+1} = 0, \delta_t = 0) \\ &= E(y_t | y_1, \dots, y_s, \delta_1 = \dots = \delta_{s+1} = 1, \delta_t = 0).\end{aligned}$$

By the property of conditional expectation and (5),

$$\begin{aligned}E[E(y_t | y_1, \dots, y_s, \delta_1 = \dots = \delta_{s+1} = 1, \delta_t = 0) | \\ y_1, \dots, y_r, \delta_1 = \dots = \delta_{r+1} = 1, \delta_t = 0] \\ &= E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_{r+1} = 1, \delta_t = 0) \\ &= E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0). \quad (6)\end{aligned}$$

This means that y_t and \tilde{y}_t have the same conditional expectation, given $y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0$. Therefore, using previously imputed values as responses in regression produces a valid estimator of $\phi_{t,r}$. Note that previously imputed values should not be used as predictors in regression, as equation (6) does not hold if some of y_1, \dots, y_r are imputed values.

Although all observed data at any time t are used for the estimation of $E(y_t)$, some but not all observed data at time $< t$ are utilized in imputation to avoid biases under nonignorable nonresponse. This is different in the ignorable nonresponse case, where typically all past observed data can be used in regression imputation.

2.3 Regression for imputation

The conditional expectations in (5) depend not only on the distribution of y_t , but also on the PSI. Even if $E(y_t | y_1, \dots, y_{t-1})$ is linear, conditional expectations in (5) are not necessarily linear, which is different from the case of $r + 1 = t$ considered in Section 2.1. An example is given by result (10) in the Appendix.

When we do not have a suitable parametric model for $\phi_{t,r}$, the nonparametric kernel regression method given in Cheng (1994) may be applied to obtain $\hat{\phi}_{t,r}$. Since the regressor (y_{1l}, \dots, y_{rl}) is multivariate when $r \geq 2$, however, kernel regression has a large variability unless the number of sampled subjects in the category defined by $\delta_{i1} = \dots = \delta_{i(r+1)} = 1$ is very large. This issue is commonly referred to as the curse of dimensionality.

Thus, we consider the following alternatives under the additional assumption that the dependence of δ_t on y_1, \dots, y_{t-1} is through a linear combination of y_1, \dots, y_{t-1} . That is,

$$P(\delta_t = 1 | y_1, \dots, y_{t-1}, \delta_1, \dots, \delta_{t-1}) = \Psi \left(\sum_{l=1}^{t-1} \gamma_l^{\delta_1, \dots, \delta_{t-1}} y_l \right), \quad (7)$$

where $\gamma_l^{\delta_1, \dots, \delta_{t-1}}$, $l = 1, \dots, t-1$, are unknown parameters depending on $\delta_1, \dots, \delta_{t-1}$ and Ψ is an unknown function with range $[0, 1]$. Under (7), it is shown in the Appendix that

$$\begin{aligned}E(y_t | z_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) \\ &= E(y_t | z_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) \\ &\quad r = 1, \dots, t-2, t = 2, \dots, T, \quad (8)\end{aligned}$$

where $z_r = \sum_{l=1}^r \gamma_{r,l} y_l$ and $\gamma_{r,l} = \gamma_l^{\delta_1, \dots, \delta_r}$ with $\delta_1 = \dots = \delta_r = 1$. Hence, to impute nonrespondents, we can condition on the linear combination z_r and use (8), instead of conditioning on y_1, \dots, y_r and using (5).

Let $\psi_{t,r}(z_r)$ be the function defined on the second line of (8). Note that $\psi_{t,r}$ is not necessary the same as $\phi_{t,r}$. If there is a strong linear relationship between y_t and y_1, \dots, y_r , then $\psi_{t,r}$ may be approximately linear so that we can fit a linear regression to obtain an estimator $\hat{\psi}_{t,r}$. In theory, this method is biased when $\psi_{t,r}$ is not linear. If $\gamma_r = (\gamma_{r,1}, \dots, \gamma_{r,r})'$ is known, then we can apply a one-dimensional kernel regression to obtain an estimator $\hat{\psi}_{t,r}$, using the one-dimensional index z_r . Since γ_r is unknown, we first need to estimate it by $\hat{\gamma}_r$ and then obtain $\hat{\psi}_{t,r}$ by applying the one-dimensional kernel regression with γ_r replaced by $\hat{\gamma}_r$. For example, the sliced inverse regression (Duan and Li 1991) can be applied to obtain $\hat{\gamma}_r$. However, this type of nonparametric method may be inefficient. If there is a strong linear relationship between y_t and y_1, \dots, y_r , we may apply linear regression to obtain $\hat{\gamma}_r$. In any case, we use y_{1l}, \dots, y_{rl} with $\delta_{i1} = \dots = \delta_{i(r+1)} = 1$ as predictors and imputed y_{it} values as responses in any type

of regression fitting. After $\hat{\psi}_{t,r}$ and $\hat{\gamma}_r = (\hat{\gamma}_{r,1}, \dots, \hat{\gamma}_{r,r})'$ are obtained, a missing y_{it} is imputed by $\tilde{y}_{it} = \hat{\psi}_{t,r}(\hat{\gamma}_{r,1}y_{i1} + \dots + \hat{\gamma}_{r,r}y_{ir})$.

We refer to the method of simply applying linear regression as the linear regression imputation method, and the method of applying kernel regression to the index z_r as the one-dimensional index kernel regression imputation method. An advantage of one-dimensional index kernel regression imputation over kernel regression imputation is that only a one-dimensional kernel regression is applied and, thus, it avoids the curse of dimensionality and has smaller variability.

These methods can also be applied to the case of $r = t - 1$ if $E(y_t | y_1, \dots, y_{t-1})$ is not linear.

In theory, estimators such as the estimated means based on kernel regression or one-dimensional index kernel regression imputation are asymptotically unbiased, but they may not be better than those based on linear regression imputation when the number of sampled subjects in each (t, r) category is not very large. The performances of the estimated means based on linear regression, kernel regression, and one-dimensional index kernel regression imputation are examined by simulation in Section 3.

2.4 Estimation

We consider the estimation of the finite population total or the mean of y_i at each fixed t , which is often the main purpose of a survey study. At any t , let $\tilde{y}_{it} = y_{it}$ when $\delta_{it} = 1$ and \tilde{y}_{it} be the imputed value using one of the methods in Section 2 when $\delta_{it} = 0$. The finite population total and the mean of y_i can be estimated by

$$\hat{Y}_t = \sum_{i \in S} w_i \tilde{y}_{it} \quad \text{and} \quad \bar{Y}_t = \sum_{i \in S} w_i \tilde{y}_{it} / \sum_{i \in S} w_i, \quad (9)$$

respectively, where w_i is the survey weight constructed such that, in the case of no nonresponse, \hat{Y}_t is an unbiased estimator of the finite population total at time t with respect to the probability sampling. The superpopulation mean of y_i can also be estimated by \bar{Y}_t . Note that $\sum_{i \in S} w_i$ is an unbiased estimator of the finite population size N and, for some simple sampling designs, it is exactly equal to N .

The survey weights should also be used in the regression fitting for imputation. Under the same conditions given in Cheng (1994), \hat{Y}_t or \bar{Y}_t based on kernel regression or one-dimensional index kernel regression imputation is consistent and asymptotically normal as the sample size increases to ∞ . The required conditions and proofs can be found in Xu (2007).

If we apply the linear regression imputation method as discussed in Section 2.3, then the resulting estimated mean at t may be asymptotically biased. This bias is small if the function $\psi_{t,r}$ can be well approximated by a linear function in the range of the data values. On the other hand, kernel or

one-dimensional index kernel regression imputation may require a much larger sample size than that for linear regression imputation. Hence, the overall performance of the estimated mean based on linear regression imputation may still be better, as indicated by the simulation results in Section 3.

2.5 Variance estimation

For assessing statistical accuracy or inference such as constructing a confidence interval for the mean of y_i at t , we need variance estimators of \hat{Y}_t or \bar{Y}_t based on imputed data. Because of the complexity of the imputation procedure, it is difficult to obtain explicit formulas for variance of \hat{Y}_t or \bar{Y}_t . The bootstrap method (Efron 1979) is then considered. A correct bootstrap can be obtained by repeating the process of imputation in each of the bootstrap samples (Shao and Sitter 1996). Let $\hat{\theta}$ be the estimator under consideration. A bootstrap procedure can be carried out as follows.

1. Draw a bootstrap sample as a simple random sample of the same size as S with replacement from the set of sampled subjects.
2. For units in the bootstrap sample, their survey weights, response indicators, and observed data from the original data set are used to form a bootstrap data set. Apply the proposed imputation procedure to the bootstrap data. Calculate the bootstrap analog $\hat{\theta}^*$ of $\hat{\theta}$.
3. Independently repeat the previous steps B times to obtain $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$. The sample variance of $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$ is the bootstrap variance estimator for $\hat{\theta}$.

In application, each $\hat{\theta}^{*b}$ can be calculated using the b^{th} bootstrap data $(y_i, \delta_i, w_i^{*b})$, $i \in S$, where $w_i^{*b} = w_i$ multiplied by the number of times unit i appears in the b^{th} bootstrap sample. Note that the same w_i^{*b} can be used for all variables of interest, not just y_i .

3. Empirical results

We study \hat{Y}_t or \bar{Y}_t in (9) based on the proposed imputation methods at each time point t . We first consider a simulation with a normal population for the y_i 's. An application to the SIRD data is presented next. To examine the performance of the proposed methods for the SIRD, a simulation with a population generated using the SIRD data is presented in the end. We have implemented the proposed imputation methods in R (R Development Core Team 2009). To fit the required nonparametric regressions, we use the R function *loess* with default settings, which fits a local polynomial surface in one or more regressor variables. The required linear regressions are easily fit in R using the

function lm . Our implementations of the proposed methods include error checking; (such as ensuring that there are sufficient points for regression fitting at each stage) which is particularly important in bootstrap and simulation settings where the imputation methods are replicated many times, and each iteration cannot be examined manually. We defaulted to an overall mean imputation in cases where there were not enough data points to fit a regression.

3.1 Simulation results from a normal population

A simulation study was conducted with normally distributed y_1, \dots, y_n , $n = 2,000$, and $T = 4$. A single imputation class and simple random sampling with replacement was considered. In the simulation, y_i 's were independently generated from the multivariate normal distribution with mean vector $(1.33, 1.94, 2.73, 3.67)$ and the covariance matrix having the AR(1) structure with correlation coefficient 0.7 and unit variance; all data at $t = 1$ were observed; missing data at $t = 2, 3, 4$ were generated according to

$$P(\delta_t = 1 | y_1, \dots, y_{t-1}, \delta_1, \dots, \delta_{t-1}) = 1 - \Phi\left(0.6\left(1 - \sum_{j=1}^{t-1} y_j \gamma_j^{\delta_1, \dots, \delta_{t-1}}\right)\right)$$

where

$$\gamma_j^{\delta_1, \dots, \delta_{t-1}} = \frac{j + (1 - \delta_j)j}{\sum_{k=1}^{t-1} [k + (1 - \delta_k)k]}, \quad j = 1, \dots, t-1,$$

and Φ is the standard normal distribution function. The unconditional probabilities of nonresponse patterns are given in Table 2.

For comparison, we included a total of nine estimators of the mean of y_t : they are sample means based on (1) the complete data (used as the gold standard); (2) respondents with adjusted weights assuming the probability of response is the same within each imputation class; (3) censoring and linear regression imputation, which first discards all observations of a subject after the first missing value to create a dataset with "monotone nonresponse" and then applies linear regression imputation as described in Paik (1997); (4) the proposed kernel regression imputation; (5) the proposed linear regression imputation; (6) the proposed one-dimensional index kernel regression imputation using the sliced inverse regression to obtain $\hat{\gamma}_r$; (7) the kernel regression imputation proposed in Xu *et al.* (2008) based on the last-value-dependent PSI; (8) the linear regression imputation based on a regression between respondents at time t and observed and imputed values at time points $1, \dots, t-1$ (treating imputed as observed); (9) the linear regression imputation based on a regression between respondents at time t and observed data from units with the same missing pattern at time points $1, \dots, t-1$.

Table 2
Probabilities of nonresponse patterns in the simulation study (Normal population)

	Pattern	Probability
Monotone	(1,0,0,0)	0.062
	(1,1,0,0)	0.043
	(1,1,1,0)	0.076
		total = 0.181
Intermittent	(1,0,0,1)	0.113
	(1,0,1,0)	0.071
	(1,0,1,1)	0.186
	(1,1,0,1)	0.124
		total = 0.494
Complete	(1,1,1,1)	0.325

Method (2) simply ignores nonrespondents and, hence, is biased and inefficient. Under the PSI assumption (1) methods (7)-(9) are also biased for $t \geq 3$, because method (7) requires the last-value-dependent assumption that is stronger than (1), method (8) treats previously imputed values as observed in regression, and method (9) requires the following condition that is not true under (1):

$$\begin{aligned} E(y_t | y_1, \dots, y_{t-1}, \delta_1 = j_1, \dots, \delta_{t-1} = j_{t-1}, \delta_t = 0) \\ = E(y_t | y_1, \dots, y_{t-1}, \delta_1 = j_1, \dots, \delta_{t-1} = j_{t-1}, \delta_t = 1) \end{aligned}$$

where (j_1, \dots, j_{t-1}) is a fixed missing pattern. Finally, as we discussed in Section 2.3, method (5) is also biased for $t \geq 3$ since linear regression is not an exactly correct model. However, methods (5), (8), and (9) may still perform well when the biases are not substantial, because the use of a simpler model and more data in regression for imputation may compensate for the loss in biased imputation. Furthermore, any assumption on the PSI may hold only approximately and it is desired to empirically study various methods in any particular application.

For the case of $r = t-1$, linear regression imputation is applied as discussed in Section 2.1. Hence, methods (3)-(6), (8)-(9) all give the same results when $t = 2$.

Table 3 reports (based on 1,000 simulation runs) the relative bias and standard deviation (SD) of the mean estimator, the mean of \widehat{SD}_{boot} , the bootstrap estimator of SD based on 200 bootstrap replications, and the coverage probability of the approximate 95% confidence interval (CI) obtained using point estimator $\pm 1.96 \times \widehat{SD}_{boot}$. The following is a summary of the results in Table 3.

1. The sample mean based on ignoring missing data is clearly biased. Although in the case of $t = 4$ its relative bias is only 3.5%, it still leads to a very low coverage probability of the confidence interval, because the SD of the estimated mean is also very small.
2. The bootstrap estimator of standard deviation performs well in all cases, even when the mean estimator is biased.
3. \bar{Y}_t based on censoring and linear regression imputation has negligible bias so that the related

confidence interval has a coverage probability close to the nominal level 95%; but it has a large SD when $t = 3$ or $t = 4$. The inefficiency of this method is obviously caused by discarding observed data from nearly 50% of sampled subjects who have intermittent nonresponse. Its performance becomes worse as t increases.

4. \bar{Y}_t based on the proposed kernel regression imputation has a relative bias between 0.0% and 0.5%, but the bias is large enough to result in a poor coverage performance of the related confidence interval at $t = 4$.

5. \bar{Y}_t based on the proposed linear regression imputation has negligible bias as well as a variance smaller than that of \bar{Y}_t based on kernel regression. The related confidence interval has a coverage probability close to the nominal level 95%.
6. \bar{Y}_t based on the proposed one-dimensional index kernel regression imputation is generally good but slightly worse than that based on the linear regression imputation.
7. \bar{Y}_t based on methods (7)-(9) has non-negligible bias when $t = 3$ or $t = 4$, which results in poor performance of the related confidence interval.

Table 3
Simulation results for mean estimation (Normal population)

Method	Quantity	$t = 2$	$t = 3$	$t = 4$
Complete data	relative bias	0%	0%	0%
	SD	0.0221	0.0223	0.0221
	\widehat{SD}_{boot}	0.0223	0.0223	0.0224
	CI coverage	94.9%	94.4%	95.4%
Respondents only	relative bias	12.8%	6.8%	3.5%
	SD	0.0282	0.0272	0.0248
	\widehat{SD}_{boot}	0.0285	0.0267	0.0252
	CI coverage	0.0%	0.0%	0.2%
Censoring and linear regression imputation	relative bias	0.0%	0.0%	-0.1%
	SD	0.0275	0.0358	0.0418
	\widehat{SD}_{boot}	0.0276	0.0354	0.0431
	CI coverage	95.1%	94.6%	95.6%
Proposed kernel regression imputation	relative bias	0.0%	0.4%	0.5%
	SD	0.0275	0.0288	0.0283
	\widehat{SD}_{boot}	0.0276	0.0288	0.0288
	CI coverage	95.1%	92.5%	88.6%
Proposed linear regression imputation	relative bias	0.0%	0.1%	0.0%
	SD	0.0275	0.0286	0.0279
	\widehat{SD}_{boot}	0.0276	0.0287	0.0293
	CI coverage	95.1%	93.8%	95.7%
Proposed 1-dimensional index kernel regression imputation	relative bias	0.0%	0.4%	0.4%
	SD	0.0275	0.0288	0.0279
	\widehat{SD}_{boot}	0.0276	0.0288	0.0288
	CI coverage	95.1%	92.5%	91.7%
Last-value-dependent kernel regression imputation	relative bias	0.6%	1.0%	0.6%
	SD	0.0284	0.0310	0.0257
	\widehat{SD}_{boot}	0.0288	0.0295	0.0263
	CI coverage	93.7%	84.2%	86.2%
Linear regression imputation treating previously imputed values as observed	relative bias	0.0%	1.6%	0.8%
	SD	0.0275	0.0261	0.0241
	\widehat{SD}_{boot}	0.0276	0.0260	0.0246
	CI coverage	95.1%	59.7%	76.0%
Linear regression imputation based on currently and previously observed data	relative bias	0.0%	1.6%	0.8%
	SD	0.0275	0.0261	0.0242
	\widehat{SD}_{boot}	0.0276	0.0261	0.0246
	CI coverage	95.1%	59.0%	76.1%

Although the kernel regression is asymptotically valid, in this simulation study the total number of subjects is 2,000 and, according to Table 2, the average numbers of data points used in kernel regression under patterns $(t, r) = (4, 1)$ and $(4, 2)$ are 238 and 152, respectively, which may not be enough for kernel regression and lead to some small biases in imputation. On the other hand, linear regression is more stable and works well with a sample size such as 152. Although linear regression imputation has a bias in theory, the bias may be small when $E(y_i | y_1, \dots, y_{t-1})$ is linear.

3.2 Application to the SIRD

The SIRD is an annual survey of about 31,000 companies potentially involved in research and development. The NSF sponsors this survey as part of a mandate requiring that NSF collect, interpret, and analyze data on scientific and engineering resources in the United States. The survey is conducted jointly by the U.S. Census Bureau and NSF. The surveyed companies are asked to provide information related to their total research and development (RD) expenditure for the calendar year of the survey. The SIRD deterministically surveys some companies each year by placing them in a certainty stratum, since they account for a large percentage of the total RD dollar investment in the U.S. The remaining companies that appear in the survey are sampled each year using a stratified probability proportionate to size (PPS) sampling design. Longitudinal measurements are available on the core of companies that are sampled with certainty and on other companies that happen to be selected each year. For the purposes of illustrating our imputation methods, we restrict attention to only those companies that were selected for the survey in each of the years 2002 through 2005 ($T = 4$), and companies that provided a response in 2002. For documentation on the SIRD and detailed statistical tables, we refer to the document titled *Research and Development in Industry: 2005*, available from <http://www.nsf.gov/statistics/nsf10319>. Additional information on the Business R&D and Innovation Survey is available online at <http://bhs.dev.econ.census.gov/bhs/brdis/> and <http://www.nsf.gov/statistics/srvyindustry/about/brdis/>.

We divide the data into two imputation classes. One class consists of all companies contained in a certainty stratum for each of the four years; the other consists of the rest of companies. Within each imputation class, the data take the form (y_i, δ_i) , $i = 1, \dots, n$, where y_{it} represents the total RD expenditure for company i at time $t = 1$ (2002), 2 (2003), 3 (2004), 4 (2005). The sample size here is $n = 2,309$ for the certainty strata class and $n = 1,039$ for the non-certainty strata class. Missingness is nonmonotone and the missing percentages for the years 2003, 2004, and 2005 were 10.4%, 14.0%, and 18.8%, for the certainty strata

class, and 15.2%, 20.7%, and 26.0% for the non-certainty strata class.

Table 4 shows the estimated totals and standard errors obtained by using the methods (2)-(9) described in the simulation study in Section 3.1. As discussed in the end of Section 2.1, in each of the proposed imputation methods we use linear regression when $r + 1 = t$. The standard errors shown in Table 4 were computed using the bootstrap method. Table 4 also displays estimated totals obtained when missing data are filled in by the values that were put in place by the Census Bureau in order to produce the officially published data tables (officially published data tables are available from http://www.nsf.gov/statistics/pubseri.cfm?seri_id=26). The method that was used by the Census Bureau to handle missing data when producing these published data tables (which we call the "current method") was ratio imputation for companies with prior year data using imputation cells formed by industry type; we refer to Bond (1994) for further details. Table 4 also presents the estimated RD totals obtained from respondents only with no weight adjustment which indicate that ignoring the missing data leads to biased estimates. Methods (3)-(9) give comparable results, which is likely due to the strong linear dependence in the data so that theoretically biased methods exhibit negligible bias. The estimated totals based on the current method are comparable to those based on the proposed methods for the certainty strata case, but are different in the non-certainty strata case. The method of censoring and linear regression has similar SD to the proposed methods because the number of data points discarded under censoring is not too large. In the certainty strata imputation class only 10% of the sample has an intermittent nonresponse pattern and the percentage of complete cases is 72%. In the non-certainty class, only 9% of the sample has an intermittent nonresponse pattern and the percentage of complete cases is 66%.

3.3 Simulation results based on the SIRD population

An additional simulation study was conducted using a population constructed from the SIRD data. The simulation was run independently for the certainty strata and non-certainty strata imputation classes. To construct the population, we begin with the SIRD data with missing values imputed using the current imputation method for the SIRD. Let δ_i be the observed response indicator vector for company i and \tilde{y}_i be the vector of either the observed or imputed values of total RD expenditures for company i over time, $i = 1, \dots, n$. For the simulation, we sample from a population based on $\{(\tilde{y}_i, \delta_i), i = 1, \dots, n\}$ as follows. We first draw a sample of size n with replacement from $\tilde{y}_1, \dots, \tilde{y}_n$, then we add independent normal random noise, with mean 0 and standard deviation 500, to each component

of each of the sampled vectors. Any resulting negative values are set to zero. We denote these simulated RD totals by y_1^*, \dots, y_n^* , where n is the same as that in Section 3.2. We denote the simulated response indicators by $\delta_1^*, \dots, \delta_n^*$. For all i and each $t = 2, 3, 4$, δ_{it}^* 's were binary random variables with

$$P(\delta_{it}^* = 1 \mid y_{i1}^*, \dots, y_{i,t-1}^*) = \frac{\exp(\beta_0^{(t)} + \beta_1^{(t)} y_{i1}^* + \dots + \beta_{t-1}^{(t)} y_{i,t-1}^*)}{1 + \exp(\beta_0^{(t)} + \beta_1^{(t)} y_{i1}^* + \dots + \beta_{t-1}^{(t)} y_{i,t-1}^*)}.$$

The coefficients $\beta_0^{(t)}, \beta_1^{(t)}, \dots, \beta_{t-1}^{(t)}$ are fixed throughout the simulation and they were obtained as the estimated coefficients from an initial fit of a logistic regression of δ_{it} on $(\tilde{y}_{i1}, \dots, \tilde{y}_{i,t-1})$ for $i = 1, \dots, n$.

Table 5 reports the simulation results for total estimators based on 1,000 runs and methods (1)-(9) described in Section 3.1, where the quantities appearing in the table are defined in Section 3.1. To compute the relative bias we obtain the true value of the total through a preliminary run of the simulation model. Several of the conclusions from the normal population simulation of Section 3.1 carry over to this setting. The following is a summary of some additional findings.

1. In contrast to the normal population simulation setting, the estimated total based on censoring and linear regression has SD that is comparable with the proposed imputation methods. This is because the number of data points discarded under censoring is small in this case. The probabilities of an intermittent response pattern are 17% and 19% for the certainty and non-certainty strata classes, respectively. In the normal population simulation these probabilities were nearly 50% as shown in Table 2.
2. All of the proposed imputation methods give relatively similar performance. As noted previously, linear regression imputation is generally biased in theory. However, the bias is small because of the strong linear dependence in data.
3. Method (7) does not have a good performance at $t \geq 3$ for the non-certainty strata case, because the last-value-dependent PSI assumption does not hold.
4. Methods (8) and (9) perform well, again due to the strong linear dependence in data. Although these methods use more observed data in regression imputation, they are comparable with the proposed linear regression method.

Table 4
RD total estimates (in thousands) from SIRD data based on years 2002 to 2005.
Bootstrap standard error (in thousands) in parentheses¹

Method	Certainty strata			Non-certainty strata		
	<i>t</i> = 2	<i>t</i> = 3	<i>t</i> = 4	<i>t</i> = 2	<i>t</i> = 3	<i>t</i> = 4
Current imputation	154,066 (-)	156,754 (-)	168,015 (-)	2,694 (-)	2,790 (-)	2,782 (-)
Respondents only with no weight adjustment	149,502 (15,907)	148,300 (16,160)	159,822 (17,149)	2,448 (172)	2,553 (193)	2,419 (207)
Respondents only with adjusted weights	166,924 (17,728)	172,419 (18,720)	196,815 (21,045)	2,887 (199)	3,219 (237)	3,269 (273)
Censoring and linear regression imputation	154,824 (15,888)	159,206 (16,394)	172,631 (17,470)	2,843 (189)	3,079 (208)	3,257 (246)
Proposed kernel regression imputation	154,824 (15,888)	159,394 (16,414)	171,633 (17,603)	2,843 (189)	2,997 (199)	3,161 (290)
Proposed linear regression imputation	154,824 (15,888)	159,198 (16,383)	172,042 (17,247)	2,843 (189)	3,043 (203)	3,302 (250)
Proposed 1-dimensional index kernel regression imputation	154,824 (15,888)	159,394 (16,414)	171,494 (17,268)	2,843 (189)	2,997 (199)	3,254 (248)
Last-value-dependent kernel regression imputation	154,688 (15,900)	158,768 (16,286)	170,606 (17,234)	2,831 (188)	2,983 (197)	3,177 (240)
Linear regression imputation treating previously imputed values as observed	154,824 (15,888)	159,401 (16,390)	172,600 (17,306)	2,843 (189)	3,098 (208)	3,257 (236)
Linear regression imputation based on currently and previously observed data	154,824 (15,888)	160,205 (16,534)	172,452 (17,209)	2,843 (189)	3,168 (233)	3,273 (254)

¹ Disclaimer: The values in Table 4 do not necessarily represent national estimates because we have made some restrictions on the data to fit our framework.

Table 5

Simulation results for total estimation (in thousands) SIRD based population

Method	Quantity	Certainty Strata			Non-Certainty Strata		
		$t = 2$	$t = 3$	$t = 4$	$t = 2$	$t = 3$	$t = 4$
Complete data	relative bias	0%	0.1%	0.1%	0.2%	0.0%	0.4%
	SD	15,541	16,045	16,947	184	203	224
	\widehat{SD}_{boot}	15,654	15,994	16,941	186	201	218
	CI coverage	94.0%	94.0%	94.3%	94.3%	93.7%	93.9%
Respondents only with adjusted weights	relative bias	5%	6.3%	11.6%	-1.1%	1.1%	-2.7%
	SD	16,870	17,858	20,032	191	220	244
	\widehat{SD}_{boot}	16,917	17,915	20,048	192	219	234
	CI coverage	94.8%	94.8%	87.3%	93.2%	94.5%	89.8%
Censoring and linear regression imputation	relative bias	0%	0.4%	0.5%	0.4%	0.1%	-0.4%
	SD	15,582	16,272	17,247	191	214	238
	\widehat{SD}_{boot}	15,654	16,145	17,195	194	214	236
	CI coverage	93.8%	93.5%	94.2%	94.8%	94.0%	93.7%
Proposed kernel regression imputation	relative bias	0%	0.2%	-0.1%	0.4%	-0.3%	-0.3%
	SD	15,582	16,130	17,098	191	205	246
	\widehat{SD}_{boot}	15,654	16,072	17,231	194	204	262
	CI coverage	93.8%	93.5%	94.2%	94.8%	93.4%	93.7%
Proposed linear regression imputation	relative bias	0%	0.2%	0.0%	0.4%	0.0%	-0.5%
	SD	15,582	16,130	16,955	191	206	229
	\widehat{SD}_{boot}	15,654	16,072	16,964	194	206	224
	CI coverage	93.8%	93.5%	94.2%	94.8%	94.0%	93.7%
Proposed 1-dimensional index kernel regression imputation	relative bias	0%	0.2%	-0.1%	0.4%	-0.3%	-0.9%
	SD	15,582	16,130	16,957	191	205	227
	\widehat{SD}_{boot}	15,654	16,072	16,965	194	204	220
	CI coverage	93.8%	93.5%	94.3%	94.8%	93.4%	93.1%
Last-value-dependent kernel regression imputation	relative bias	0%	0.1%	-0.3%	0.0%	-0.7%	-0.7%
	SD	15,565	16,019	16,990	184	204	242
	\widehat{SD}_{boot}	15,635	16,003	16,983	187	202	230
	CI coverage	93.8%	93.7%	94.0%	93.9%	92.7%	91.1%
Linear regression imputation treating previously imputed values as observed	relative bias	0%	0.2%	0.0%	0.4%	0.6%	-0.6%
	SD	15,582	16,120	16,952	191	210	231
	\widehat{SD}_{boot}	15,654	16,065	16,954	194	210	225
	CI coverage	93.8%	93.6%	94.3%	94.8%	93.8%	92.8%
Linear regression imputation based on currently and previously observed data	relative bias	0%	0.2%	0.0%	0.4%	0.6%	-0.6%
	SD	15,582	16,117	16,945	191	213	241
	\widehat{SD}_{boot}	15,654	16,062	16,954	194	211	254
	CI coverage	93.8%	93.5%	94.3%	94.8%	93.6%	93.7%

4. Concluding remarks

We consider a longitudinal study variable having non-monotone nonresponse. Under the assumption that the PSI depends on past observed or unobserved values of the study variable, we propose several imputation methods that lead to unbiased or nearly unbiased estimators of the total or mean of the study variable at a given time point. Our methods do

not require any parametric model on the joint distribution of the variables across time points or the PSI. They are based on regression models under different nonresponse patterns derived from the past-data-dependent PSI. Three regression methods are adopted, linear regression, kernel regression, and one-dimensional index kernel regression. The imputation method based on the kernel type regression is asymptotically valid, but it requires a large number of

observations in each nonresponse pattern. The imputation method based on linear regression is asymptotically biased when the linear relationship does not hold, but it is more stable and, therefore, it may still out-perform methods based on kernel regression.

The method of censoring, which discards all observed data from a subject after its first missing value, may work well when the number of data discarded is small; otherwise it may be very inefficient especially when T is large. For the SIRD data analysis in Sections 3.2-3.3, censoring is comparable with the proposed linear regression imputation method. However, the results are based on four years of data only and censoring may lead to inefficient estimators when more years of data are considered. In applications, it may be a good idea to compare estimators based on censoring with those based on the proposed methods.

Estimators based on the linear regression imputation methods (8) and (9) described in Section 3.1 are asymptotically biased in general. Although they perform well in the simulation study based on the SIRD population, they have poor performance under the simulation setting in Section 3.1, while the proposed linear regression imputation performs well.

The results in Section 2 can be extended to the situation where each sample unit has an observed covariate \mathbf{x}_t at time t without missing values. Assumption (1) may be modified to include covariates:

$$P(\delta_t = 1 \mid \mathbf{y}, \mathbf{X}, \delta_1, \dots, \delta_{t-1}, \delta_{t+1}, \dots, \delta_T) \\ = P(\delta_t = 1 \mid y_t, \dots, y_{t-1}, \mathbf{X}, \delta_1, \dots, \delta_{t-1}), \quad t = 2, \dots, T,$$

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$. Missing components of \mathbf{y}_t can be imputed using one of the procedures in Sections 2.1-2.3 with (y_{i1}, \dots, y_{iT}) replaced by $(y_{i1}, \dots, y_{iT}, \mathbf{X}_i)$. After all missing values are imputed, we can also estimate the relationship between \mathbf{y} and \mathbf{X} using some popular approaches such as the generalized estimation equation approach. Some details can be found in Xu (2007).

It is implicitly assumed throughout the paper that the y -values are continuous variables with no restriction. When y -values have a particular order or are integer valued, the proposed regression imputation methods are clearly not suitable. New methods for these situations have to be developed.

Acknowledgements

We thank Katherine Jenny Thompson and David L. Kinyon, both of the U.S. Census Bureau, as well as two referees and the associate editor for providing many helpful comments on the paper. The research was partially supported by an NSF grant. This article is released to inform interested parties of ongoing research and to encourage

discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Appendix

Proof of (2) - (3). Let $L(\xi)$ denote the distribution of ξ and $L(\xi \mid \zeta)$ denote the conditional distribution of ξ given ζ . Let $\mathbf{y}_t = (y_1, \dots, y_t)$ and $\delta_t = (\delta_1, \dots, \delta_t)$. Then, both (2) and (3) follow from $L(y_t \mid \mathbf{y}_{t-1}, \delta_t) = L(y_t \mid \mathbf{y}_{t-1}, \delta_{t-1}) = L(\mathbf{y}_t, \delta_{t-1}) / L(\mathbf{y}_{t-1}, \delta_{t-1}) = [L(\delta_{t-1} \mid \mathbf{y}_t, \delta_{t-2}) / L(\delta_{t-1} \mid \mathbf{y}_{t-1}, \delta_{t-2})] L(y_t \mid \mathbf{y}_{t-1}, \delta_{t-2}) = L(y_t \mid \mathbf{y}_{t-1}, \delta_{t-2}) = L(y_t \mid \mathbf{y}_{t-1}, \delta_{t-3}) = \dots = L(y_t \mid \mathbf{y}_{t-1})$, where the first and third equalities follow from assumption (1).

Proof of (5). Using the same notation as in the proof of (2) and letting $\Delta_r = 1$ be the indicator of $\delta_1 = \dots = \delta_r = 1$, we have $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = [L(\delta_{r+1} = 0 \mid y_t, \mathbf{y}_r, \Delta_r = 1, \delta_t = 0) / L(\delta_{r+1} = 0 \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)] L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$, which is equal to $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$ by (1). Similarly, we can show that $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) = L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$. Hence, $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 1, \delta_t = 0)$ and result (5) follows.

An example in which (4) does not hold. To show that (4) does not hold in general, we only need to give a counterexample. Consider $T = 3$. Let (y_1, y_2, y_3) be jointly normal with $E(y_t) = 0$, $\text{var}(y_t) = 1$, $t = 1, 2, 3$, $\text{cov}(y_1, y_2) = \text{cov}(y_1, y_3) = \rho$, and $\text{cov}(y_2, y_3) = \rho^2$, where $\rho \neq 0$ is a parameter. Suppose that y_1 is always observed and $P(\delta_t = 0 \mid y_{t-1}) = \Phi(a_{t-1} + b_{t-1} y_{t-1})$, $t = 2, 3$, where a_t and b_t are parameters, Φ is the cumulative distribution function of the standard normal distribution. Then, $E(y_3 \mid y_2, y_1) = \rho y_2$, $E(y_2 \mid y_1) = \rho y_1$, and $E(y_3 \mid y_1) = \rho^2 y_1$. Note that

$$\begin{aligned} E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) &= E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = 1) \\ &= E(y_3 \mid y_1, \delta_3 = 0) \\ &= \int y_3 L(y_3 \mid y_1, \delta_3 = 0) dy_3 \\ &= \int y_3 \int L(y_3 \mid y_1, y_2, \delta_3 = 0) L(y_2 \mid y_1, \delta_3 = 0) dy_2 dy_3 \\ &= \iint y_3 L(y_3 \mid y_1, y_2) L(y_2 \mid y_1, \delta_3 = 0) dy_2 dy_3 \\ &= \int \left(\int y_3 L(y_3 \mid y_2) dy_3 \right) L(y_2 \mid y_1, \delta_3 = 0) dy_2 \\ &= \rho \int y_2 L(y_2 \mid y_1, \delta_3 = 0) dy_2 \\ &= \frac{\rho \int y_2 P(\delta_3 = 0 \mid y_2) L(y_2 \mid y_1) dy_2}{\int P(\delta_3 = 0 \mid y_2) L(y_2 \mid y_1) dy_2} \\ &= \frac{\rho \int y_2 \Phi(a_2 + b_2 y_2) L(y_2 \mid y_1) dy_2}{\int \Phi(a_2 + b_2 y_2) L(y_2 \mid y_1) dy_2}, \end{aligned}$$

where the first equality holds because y_1 is always observed, the second equality holds because under (1), δ_2 and y_3 are independent given y_1 . The denominator of the previous expression is equal to

$$h(y_1) = \Phi\left(\frac{a_2 + b_2 \rho y_1}{\sqrt{1 + b_2^2(1 - \rho^2)}}\right).$$

Using integration by parts, we obtain that

$$\begin{aligned} g(y_1) &= \int (y_2 - \rho y_1) \Phi(a_2 + b_2 y_2) L(y_2 | y_1) dy_2 \\ &= b_2(1 - \rho^2) \int \Phi(a_2 + b_2 y_2) L(y_2 | y_1) dy_2 \\ &= \frac{b_2^2(1 - \rho^2)}{2\pi\sqrt{1 - \rho^2}} \int \exp\left\{-\frac{(a_2 + b_2 y_2)^2}{2} - \frac{(y_2 - \rho y_1)^2}{2(1 - \rho^2)}\right\} dy_2 \\ &= \frac{b_2(1 - \rho^2)}{2\pi[1 + b_2^2(1 - \rho^2)]} \exp\left\{-\frac{(a_2 + b_2 \rho y_1)^2}{2[1 + b_2^2(1 - \rho^2)]}\right\}. \end{aligned}$$

Thus,

$$E(y_3 | y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) = \rho^2 y_1 + \rho \frac{g(y_1)}{h(y_1)}. \quad (10)$$

However,

$$\begin{aligned} E(y_3 | y_1, \delta_1 = \delta_2 = 1) &= E(y_3 | y_1, \delta_1 = 1) \\ &= E(y_3 | y_1) = \rho^2 y_1. \end{aligned}$$

This shows that (4) does not hold in this special case.

Proof of (8). Using the notation in the proof of (2)-(3) and writing the $(t-2)$ -dimensional vector $(y_1, \dots, y_{r-1}, y_{r+1}, \dots, y_{t-1})$ as $\mathbf{u}_{t,r}$, we obtain that

$$\begin{aligned} L(\delta_{r+1} = 1 | y_t, z_r, \Delta_r = 1, \delta_t = 0) &= \int L(\delta_{r+1} = 1 | y_t, z_r, \mathbf{u}_{t,r}, \Delta_r = 1, \delta_t = 0) \\ &\quad L(\mathbf{u}_{t,r} | y_t, z_r, \Delta_r = 1, \delta_t = 0) d\mathbf{u}_{t,r} \\ &= \int L(\delta_{r+1} = 1 | y_1, \dots, y_r, \Delta_r = 1) \\ &\quad L(\mathbf{u}_{t,r} | y_t, z_r, \Delta_r = 1, \delta_t = 0) d\mathbf{u}_{t,r} \\ &= \int L(\delta_{r+1} = 1 | z_r, \Delta_r = 1) \\ &\quad L(\mathbf{u}_{t,r} | y_t, z_r, \Delta_r = 1, \delta_t = 0) d\mathbf{u}_{t,r} \\ &= L(\delta_{r+1} = 1 | z_r, \Delta_r = 1) \\ &\quad \int L(\mathbf{u}_{t,r} | y_t, z_r, \Delta_r = 1, \delta_t = 0) d\mathbf{u}_{t,r} \\ &= L(\delta_{r+1} = 1 | z_r, \Delta_r = 1), \end{aligned}$$

where the second equality follows from assumption (1) and the fact that there is a one-to-one function between $(z_r, \mathbf{u}_{t,r})$ and (y_1, \dots, y_{t-1}) , and the third equality follows from assumption (7). Similarly, $L(\delta_{r+1} = 1 | z_r, \Delta_r = 1, \delta_t = 0) = L(\delta_{r+1} = 1 | z_r, \Delta_r = 1)$ and, hence, $L(\delta_{r+1} = 1 | y_t, z_r, \Delta_r = 1, \delta_t = 0) = L(\delta_{r+1} = 1 | z_r, \Delta_r = 1, \delta_t = 0)$. Then,

$$\begin{aligned} L(y_t | z_r, \Delta_{r+1} = 1, \delta_t = 0) &= \frac{L(y_t, z_r, \Delta_{r+1} = 1, \delta_t = 0)}{L(z_r, \Delta_{r+1} = 1, \delta_t = 0)} \\ &= \frac{L(\delta_{r+1} = 1 | y_t, z_r, \Delta_r = 1, \delta_t = 0)}{L(\delta_{r+1} = 1 | z_r, \Delta_r = 1, \delta_t = 0)} \frac{L(y_t, z_r, \Delta_r = 1, \delta_t = 0)}{L(z_r, \Delta_r = 1, \delta_t = 0)} \\ &= L(y_t | z_r, \Delta_r = 1, \delta_t = 0). \end{aligned}$$

Similarly, $L(y_t | z_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t | z_r, \Delta_r = 1, \delta_t = 0)$. Hence, $L(y_t | z_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t | z_r, \Delta_{r+1} = 1, \delta_t = 0)$ and result (8) follows.

References

- Bond, D. (1994). An evaluation of imputation methods for the Survey of Industrial Research and Development. *U.S. Bureau of the Census, Economic Statistical Methods and Programming Division Report Series*, 9404. Washington, DC.
- Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89, 81-87.
- Diggle, P., and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 43, 49-93.
- Duan, N., and Li, K.C. (1991). Sliced regression: A link-free regression method. *The Annals of Statistics*, 19, 505-530.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.
- Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1, 1-16.
- Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.
- Little, R.J., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, second edition. New York: John Wiley & Sons, Inc.
- National Science Foundation, Division of Science Resources Statistics (2010). *Research and Development in Industry: 2005*. Detailed Statistical Tables. Available from <http://www.nsf.gov/statistics/nsf10319/>.
- Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92, 1320-1329.

- R Development Core Team (2009). A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0.
- Robins, J.M., and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122-129.
- Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.
- Troxel, A.B., Harrington, D.P. and Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, 47, 425-438.
- Troxel, A.B., Lipsitz, S.R. and Harrington, D.P. (1998). Marginal models for the analysis of longitudinal measurements with non-ignorable non-monotone missing data. *Biometrika*, 85, 661-672.
- Vansteelandt, S., Rotnitzky, A. and Robins, J.M. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94, 841-860.
- Xu, J. (2007). Methods for intermittent missing responses in longitudinal data. Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.
- Xu, J., Shao, J., Palta, M. and Wang, L. (2008). Imputation for nonmonotone last-value-dependent nonrespondents in longitudinal surveys. *Survey Methodology*, 34, 2, 153-162.

Some theory for propensity-score-adjustment estimators in survey sampling

Jae Kwang Kim and Minsun Kim Riddles¹

Abstract

The propensity-scoring-adjustment approach is commonly used to handle selection bias in survey sampling applications, including unit nonresponse and undercoverage. The propensity score is computed using auxiliary variables observed throughout the sample. We discuss some asymptotic properties of propensity-score-adjusted estimators and derive optimal estimators based on a regression model for the finite population. An optimal propensity-score-adjusted estimator can be implemented using an augmented propensity model. Variance estimation is discussed and the results from two simulation studies are presented.

Key Words: Calibration; Missing data; Nonresponse; Weighting.

1. Introduction

Consider a finite population of size N , where N is known. For each unit i , y_i is the study variable and \mathbf{x}_i is the q -dimensional vector of auxiliary variables. The parameter of interest is the finite population mean of the study variable, $\theta = N^{-1} \sum_{i=1}^N y_i$. The finite population $\mathcal{F}_N = \{(\mathbf{x}'_1, y_1), (\mathbf{x}'_2, y_2), \dots, (\mathbf{x}'_N, y_N)\}$ is assumed to be a random sample of size N from a superpopulation distribution $F(\mathbf{x}, y)$. Suppose a sample of size n is drawn from the finite population according to a probability sampling design. Let $w_i = \pi_i^{-1}$ be the design weight, where π_i is the first-order inclusion probability of unit i obtained from the probability sampling design. Under complete response, the finite population mean can be estimated by the Horvitz-Thompson (HT) estimator, $\hat{\theta}_{HT} = N^{-1} \sum_{i \in A} w_i y_i$, where A is the set of indices appearing in the sample.

In the presence of missing data, the HT estimator $\hat{\theta}_{HT}$ cannot be computed. Let r be the response indicator variable that takes the value one if y is observed and takes the value zero otherwise. Conceptually, as discussed by Fay (1992), Shao and Steel (1999), and Kim and Rao (2009), the response indicator can be extended to the entire population as $\mathcal{R}_N = \{r_1, r_2, \dots, r_N\}$, where r_i is a realization of the random variable r . In this case, the complete-case (CC) estimator $\hat{\theta}_{CC} = \sum_{i \in A} w_i r_i y_i / \sum_{i \in A} w_i r_i$ converges in probability to $E(Y | r = 1)$. Unless the response mechanism is missing completely at random in the sense that $E(Y | r = 1) = E(Y)$, the CC estimator is biased. To correct for the bias of the CC estimator, if the response probability

$$p(\mathbf{x}, y) = \Pr(r = 1 | \mathbf{x}, y) \quad (1)$$

is known, then the weighted CC estimator $\hat{\theta}_{WCC} = N^{-1} \sum_{i \in A} w_i r_i y_i / p(\mathbf{x}_i, y_i)$ can be used to estimate θ . Note that $\hat{\theta}_{WCC}$ is unbiased because $E\{\sum_{i \in A} w_i r_i y_i / p(\mathbf{x}_i, y_i) | \mathcal{F}_N\} = E\{\sum_{i=1}^N r_i y_i / p(\mathbf{x}_i, y_i) | \mathcal{F}_N\} = \sum_{i=1}^N y_i$.

If the response probability (1) is unknown, one can postulate a parametric model for the response probability $p(\mathbf{x}, y; \phi)$ indexed by $\phi \in \Omega$ such that $p(\mathbf{x}, y) = p(\mathbf{x}, y; \phi_0)$ for some $\phi_0 \in \Omega$. We assume that there exists a $1/n$ -consistent estimator $\hat{\phi}$ of ϕ_0 such that

$$\sqrt{n}(\hat{\phi} - \phi_0) = O_p(1), \quad (2)$$

where $g_n = O_p(1)$ indicates g_n is bounded in probability. Using $\hat{\phi}$, we can obtain the estimated response probability by $\hat{p}_i = p(\mathbf{x}_i, y_i; \hat{\phi})$, which is often called the propensity score (Rosenbaum and Rubin 1983). The propensity-score-adjusted (PSA) estimator can be constructed as

$$\hat{\theta}_{PSA} = \frac{1}{N} \sum_{i \in A} w_i \frac{r_i}{\hat{p}_i} y_i. \quad (3)$$

The PSA estimator (3) is widely used. Many surveys use the PSA estimator to reduce nonresponse bias (Fuller, Loughin and Baker 1994; Rizzo, Kalton and Brick 1996). Rosenbaum and Rubin (1983) and Rosenbaum (1987) proposed using the PSA approach to estimate the treatment effects in observational studies. Little (1988) reviewed the PSA methods for handling unit nonresponse in survey sampling. Duncan and Stasny (2001) used the PSA approach to control coverage bias in telephone surveys. Folsom (1991) and Iannacchione, Milne and Folsom (1991) used a logistic regression model for the response probability estimation. Lee (2006) applied the PSA method to a volunteer panel web survey. Durrant and Skinner (2006) used the PSA approach to address measurement error.

Despite the popularity of PSA estimators, asymptotic properties of PSA estimators have not received much attention in survey sampling literature. Kim and Kim (2007) used a Taylor expansion to obtain the asymptotic mean and variance of PSA estimators and discussed variance estimation. Da Silva and Opsomer (2006) and Da Silva and

Opsomer (2009) considered nonparametric methods to obtain PSA estimators.

In this paper, we discuss optimal PSA estimators in the class of PSA estimators of the form (3) that use a \sqrt{n} -consistent estimator $\hat{\phi}$. Such estimators are asymptotically unbiased for θ . Finding minimum variance PSA estimators among this particular class of PSA estimators is a topic of major interest in this paper.

Section 2 presents the main results. An optimal PSA estimator using an augmented propensity score model is proposed in Section 3. In Section 4, variance estimation of the proposed estimator is discussed. Results from two simulation studies can be found in Section 5 and concluding remarks are made in Section 6.

2. Main results

In this section, we discuss some asymptotic properties of PSA estimators. We assume that the response mechanism does not depend on y . Thus, we assume that

$$\Pr(r = 1 | \mathbf{x}, y) = \Pr(r = 1 | \mathbf{x}) = p(\mathbf{x}; \phi_0) \quad (4)$$

for some unknown vector ϕ_0 . The first equality implies that the data are missing-at-random (MAR), as we always observe \mathbf{x} in the sample. Note that the MAR condition is assumed in the population model. In the second equality, we further assume that the response mechanism is known up to an unknown parameter ϕ_0 . The response mechanism is slightly different from that of Kim and Kim (2007), where the response mechanism is assumed to be under the classical two-phase sampling setup and depends on the realized sample:

$$\Pr(r = 1 | \mathbf{x}, y, I = 1) = \Pr(r = 1 | \mathbf{x}, I = 1) = p(\mathbf{x}; \phi_A^0). \quad (5)$$

Here, I is the sampling indicator function defined throughout the population. That is, $I_i = 1$ if $i \in A$ and $I_i = 0$ otherwise. Unless the sampling design is non-informative in the sense that the sample selection probabilities are correlated with the response indicator even after conditioning on auxiliary variables (Pfeffermann, Krieger and Rinott 1998), the two response mechanisms, (4) and (5), are different. In survey sampling, assumption (4) is more appropriate because an individual's decision on whether or not to respond to a survey is at his or her own discretion. Here, the response indicator variable r_i is defined throughout the population, as discussed in Section 1.

We consider a class of \sqrt{n} -consistent estimators of ϕ_0 in (4). In particular, we consider a class of estimators which can be written as a solution to

$$\hat{\mathbf{U}}_h(\phi) \equiv \sum_{i \in A} w_i \{r_i - p_i(\phi)\} \mathbf{h}_i(\phi) = \mathbf{0}, \quad (6)$$

where $p_i(\phi) = p(\mathbf{x}_i; \phi)$ for some function $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi)$, a smooth function of \mathbf{x}_i and parameter ϕ . Thus, the solution to (6) can be written as $\hat{\phi}_h$, which depends on the choice of $\mathbf{h}_i(\phi)$. Any solution $\hat{\phi}_h$ to (6) is consistent for ϕ_0 in (4) because $E\{\hat{\mathbf{U}}_h(\phi_0) | \mathcal{F}_N\} = E[\sum_{i=1}^N \{r_i - p_i(\phi_0)\} \mathbf{h}_i(\phi_0) | \mathcal{F}_N]$ is zero under the response mechanism in (4). If we drop the sampling weights w_i in (6), the estimated parameter $\hat{\phi}_h$ is consistent for ϕ_A^0 in (5) and the resulting PSA estimator is consistent only when the sampling design is non-informative. The PSA estimators obtained from (6) using the sampling weights are consistent regardless of whether the sampling design is non-informative or not. According to Chamberlain (1987), any \sqrt{n} -consistent estimator of ϕ_0 in (4) can be written as a solution to (6). Thus, the choice of $\mathbf{h}_i(\phi)$ in (6) determines the efficiency of the resulting PSA estimator.

Let $\hat{\theta}_{\text{PSA},h}$ be the PSA estimator in (3) using $\hat{p}_i = p_i(\hat{\phi}_h)$ with $\hat{\phi}_h$ being the solution to (6). To discuss the asymptotic properties of $\hat{\theta}_{\text{PSA},h}$, assume a sequence of finite populations and samples, as in Isaki and Fuller (1982), such that $\sum_{i \in A} w_i \mathbf{u}_i - \sum_{i=1}^N \mathbf{u}_i = O_p(n^{-1/2}N)$ for any population characteristics \mathbf{u}_i with bounded fourth moments. We also assume that the sampling weights are uniformly bounded. That is, $K_1 < N^{-1}nw_i < K_2$ for all i uniformly in n , where K_1 and K_2 are fixed constants. In addition, we assume the following regularity conditions:

- [C1] The response mechanism satisfies (4), where $p(\mathbf{x}; \phi)$ is continuous in ϕ with continuous first and second derivatives in an open set containing ϕ_0 . The responses are independent in the sense that $\text{Cov}(r_i, r_j | \mathbf{x}) = 0$ for $i \neq j$. Also, $p(\mathbf{x}_i; \phi) > c$ for all i for some fixed constant $c > 0$.
- [C2] The solution to (6) exists and is unique almost everywhere. The function $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi)$ in (6) has a bounded fourth moment. Furthermore, the partial derivative $\partial\{\hat{\mathbf{U}}_h(\phi)\}/\partial\phi$ is nonsingular for all n .
- [C3] The estimating function $\hat{\mathbf{U}}_h(\phi)$ in (6) converges in probability to $\mathbf{U}_h(\phi) = \sum_{i=1}^N \{r_i - p_i(\phi)\} \mathbf{h}_i(\phi)$ uniformly in ϕ . Furthermore, the partial derivative $\partial\{\hat{\mathbf{U}}_h(\phi)\}/\partial\phi$ converges in probability to $\partial\{\mathbf{U}_h(\phi)\}/\partial\phi$ uniformly in ϕ . The solution ϕ_N to $\mathbf{U}_h(\phi) = \mathbf{0}$ satisfies $N^{1/2}(\phi_N - \phi_0) = O_p(1)$ under the response mechanism.

Condition [C1] states the regularity conditions for the response mechanism. Condition [C2] is the regularity condition for the solution $\hat{\phi}_h$ to (6). In Condition [C3], some regularity conditions are imposed on the estimating function $\hat{\mathbf{U}}_h(\phi)$ itself. By [C2] and [C3], we can establish the n -consistency (2) of $\hat{\phi}_h$.

Now, the following theorem deals with some asymptotic properties of the PSA estimator $\hat{\theta}_{\text{PSA},h}$.

Theorem 1 If conditions [C1] - [C3] hold, then under the joint distribution of the sampling mechanism and the response mechanism, the PSA estimator $\hat{\theta}_{\text{PSA},h}$ satisfies

$$\sqrt{n}(\hat{\theta}_{\text{PSA},h} - \tilde{\theta}_{\text{PSA},h}) = o_p(1), \quad (7)$$

where

$$\tilde{\theta}_{\text{PSA},h} = \frac{1}{N} \sum_{i \in A} w_i \left\{ p_i \mathbf{h}'_i \gamma^* + \frac{r_i}{p_i} (y_i - p_i \mathbf{h}'_i \gamma^*) \right\}, \quad (8)$$

$\gamma^* = (\sum_{i=1}^N r_i \mathbf{z}_i p_i \mathbf{h}'_i)^{-1} (\sum_{i=1}^N r_i \mathbf{z}_i y_i)$, $p_i = p(\mathbf{x}_i; \phi_0)$, $\mathbf{z}_i = \partial \{p^{-1}(\mathbf{x}_i; \phi_0)\} / \partial \phi$, and $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i; \phi_0)$. Moreover, if the finite population is a random sample from a superpopulation model, then

$$V(\tilde{\theta}_{\text{PSA},h}) \geq V_l \equiv V(\hat{\theta}_{\text{HT}}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) V(Y | \mathbf{x}_i) \right\}. \quad (9)$$

The equality in (9) holds when $\hat{\phi}_h$ satisfies

$$\sum_{i \in A} w_i \left\{ \frac{r_i}{p(\mathbf{x}_i; \hat{\phi}_h)} - 1 \right\} E(Y | \mathbf{x}_i) = 0, \quad (10)$$

where $E(Y | \mathbf{x}_i)$ is the conditional expectation under the superpopulation model.

Proof. Given $p_i(\phi) = p(\mathbf{x}_i; \phi)$ and $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi)$, define

$$\hat{\theta}(\phi, \gamma) = N^{-1} \sum_{i \in A} w_i \left[p_i(\phi) \mathbf{h}'_i(\phi) \gamma + \frac{r_i}{p_i(\phi)} \{y_i - p_i(\phi) \mathbf{h}'_i(\phi) \gamma\} \right].$$

Since $\hat{\phi}_h$ satisfies (6), we have $\hat{\theta}_{\text{PSA}} = \hat{\theta}(\hat{\phi}_h, \gamma)$ for any choice of γ . We now want to find a particular choice of γ , say γ^* , such that

$$\hat{\theta}(\hat{\phi}_h, \gamma^*) = \hat{\theta}(\phi_0, \gamma^*) + o_p(n^{-1/2}). \quad (11)$$

As $\hat{\phi}_h$ converges in probability to ϕ_0 , the asymptotic equivalence (11) holds if

$$E \left\{ \frac{\partial}{\partial \phi} \hat{\theta}(\phi, \gamma^*) \mid \phi = \phi_0 \right\} = \mathbf{0}, \quad (12)$$

using the theory of Randles (1982). Condition (12) holds if $\gamma^* = \gamma^*_h$, where γ^*_h is defined in (8). Thus, (11) reduces to

$$\hat{\theta}_{\text{PSA},h} = \frac{1}{N} \sum_{i \in A} w_i \left\{ p_i \mathbf{h}'_i \gamma^*_h + \frac{r_i}{p_i} (y_i - p_i \mathbf{h}'_i \gamma^*_h) \right\} + o_p(n^{-1/2}), \quad (13)$$

which proves (7). The variance of $\tilde{\theta}_{\text{PSA},h}$ can be derived as

$$\begin{aligned} V(\tilde{\theta}_{\text{PSA},h}) &= V(\hat{\theta}_{\text{HT}}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) (y_i - p_i \mathbf{h}'_i \gamma^*)^2 \right\} \\ &= V(\hat{\theta}_{\text{HT}}) + \frac{1}{N^2} E \left[\sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) \{y_i - E(Y | \mathbf{x}_i) + E(Y | \mathbf{x}_i) - p_i \mathbf{h}'_i \gamma^*\}^2 \right] \\ &= V(\hat{\theta}_{\text{HT}}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) V(Y | \mathbf{x}_i) \right\} \\ &\quad + \frac{1}{N^2} E \left[\sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) \{E(Y | \mathbf{x}_i) - p_i \mathbf{h}'_i \gamma^*\}^2 \right], \quad (14) \end{aligned}$$

where the last equality follows because y_i is conditionally independent of $E(Y | \mathbf{x}_i) - p_i \mathbf{h}'_i \gamma^*$, conditioning on \mathbf{x}_i . Since the last term in (14) is non-negative, the inequality in (9) is established. Furthermore, if $E(Y | \mathbf{x}_i) = p_i \mathbf{h}'_i \alpha$ for some α , then (10) holds and $E(\gamma^*_h | \mathbf{x}_i) = \alpha$, by the definition of γ^*_h . Thus, $E(Y | \mathbf{x}_i) - p_i \mathbf{h}'_i \gamma^*_h = -p_i \mathbf{h}'_i \{\gamma^*_h - E(\gamma^*_h | \mathbf{x}_i)\} = o_p(1)$, implying that the last term in (14) is negligible.

In (9), V_l is the lower bound of the asymptotic variance of PSA estimators of the form (3) satisfying (6). Any PSA estimator that has the asymptotic variance V_l in (9) is optimal in the sense that it achieves the lower bound of the asymptotic variance among the class of PSA estimators with $\hat{\phi}$ satisfying (2). The asymptotic variance of optimal PSA estimators of θ is equal to V_l in (9). The PSA estimator using the maximum likelihood estimator of ϕ_0 does not necessarily achieve the lower bound of the asymptotic variance.

Condition (10) provides a way of constructing an optimal PSA estimator. First, we need an assumption for $E(Y | \mathbf{x})$, which is often called the outcome regression model. If the outcome regression model is a linear regression model of the form $E(Y | \mathbf{x}) = \beta_0 + \beta'_1 \mathbf{x}$, an optimal PSA estimator of θ can be obtained by solving

$$\sum_{i \in A} w_i \frac{r_i}{p_i(\phi)} (1, \mathbf{x}_i) = \sum_{i \in A} w_i (1, \mathbf{x}_i). \quad (15)$$

Condition (15) is appealing because it says that the PSA estimator applied to $y = a + \mathbf{b}'\mathbf{x}$ leads to the original HT estimator. Condition (15) is called the calibration condition in survey sampling. The calibration condition applied to \mathbf{x} makes full use of the information contained in it if the study variable is well approximated by a linear function of \mathbf{x} . Condition (15) was also used in Nevo (2003) and Kott (2006) under the linear regression model.

If we explicitly use a regression model for $E(Y | \mathbf{x})$, it is possible to construct an estimator that has asymptotic variance (9) and is not necessarily a PSA estimator. For example, if we assume that

$$E(Y | \mathbf{x}) = m(\mathbf{x}; \beta_0) \quad (16)$$

for some function $m(\mathbf{x}; \cdot)$ known up to β_0 , we can use the model (16) directly to construct an optimal estimator of the form

$$\hat{\theta}_{\text{opt}} = \frac{1}{N} \sum_{i \in A} w_i \left[m(\mathbf{x}_i; \hat{\beta}) + \frac{r_i}{p_i(\hat{\phi})} \{y_i - m(\mathbf{x}_i; \hat{\beta})\} \right], \quad (17)$$

where $\hat{\beta}$ is a \sqrt{n} -consistent estimator of β_0 in the superpopulation model (16) and $\hat{\phi}$ is a \sqrt{n} -consistent estimator of ϕ_0 computed by (6). The following theorem shows that the optimal estimator (17) achieves the lower bound in (9).

Theorem 2 Let the conditions of Theorem 1 hold. Assume that $\hat{\beta}$ satisfies $\hat{\beta} = \beta_0 + O_p(n^{-1/2})$. Assume that, in the superpopulation model (16), $m(\mathbf{x}; \beta)$ has continuous first-order partial derivatives in an open set containing β_0 . Under the joint distribution of the sampling mechanism, the response mechanism, and the superpopulation model (16), the estimator $\hat{\theta}_{\text{opt}}$ in (17) satisfies

$$\sqrt{n}(\hat{\theta}_{\text{opt}} - \tilde{\theta}_{\text{opt}}^*) = o_p(1),$$

where

$$\tilde{\theta}_{\text{opt}}^* = N^{-1} \sum_{i \in A} w_i \left[m(\mathbf{x}_i; \beta_0) + \frac{r_i}{p_i} \{y_i - m(\mathbf{x}_i; \beta_0)\} \right],$$

$p_i = p_i(\phi_0)$, and $V(\tilde{\theta}_{\text{opt}}^*)$ is equal to V_l in (9).

Proof. Define $\hat{\theta}_{\text{opt}}(\beta, \phi) = N^{-1} \sum_{i \in A} w_i [m(\mathbf{x}_i; \beta) + r_i p_i^{-1}(\phi) \{y_i - m(\mathbf{x}_i; \beta)\}]$. Note that $\hat{\theta}_{\text{opt}}$ in (17) can be written as $\hat{\theta}_{\text{opt}} = \hat{\theta}_{\text{opt}}(\hat{\beta}, \hat{\phi})$. Since

$$\frac{\partial}{\partial \beta} \hat{\theta}_{\text{opt}}(\beta, \phi) = \frac{1}{N} \sum_{i \in A} w_i \left\{ \tilde{m}(\mathbf{x}_i; \beta) - \frac{r_i}{p_i(\phi)} \tilde{m}(\mathbf{x}_i; \beta) \right\},$$

where $\tilde{m}(\mathbf{x}_i; \beta) = \partial m(\mathbf{x}_i; \beta) / \partial \beta$, and

$$\frac{\partial}{\partial \phi} \hat{\theta}_{\text{opt}}(\beta, \phi) = \frac{1}{N} \sum_{i \in A} w_i r_i \mathbf{z}_i(\phi) \{y_i - m(\mathbf{x}_i; \beta)\},$$

where $\mathbf{z}_i(\phi) = \partial \{p_i^{-1}(\phi)\} / \partial \phi$, we have $E[\partial \{\hat{\theta}_{\text{opt}}(\beta, \phi)\} / \partial(\beta, \phi) | \beta = \beta_0, \phi = \phi_0] = \mathbf{0}$ and the condition of Randles (1982) is satisfied. Thus,

$$\hat{\theta}_{\text{opt}}(\hat{\beta}, \hat{\phi}) = \hat{\theta}_{\text{opt}}(\beta_0, \phi_0) + o_p(n^{-1/2}) = \tilde{\theta}_{\text{opt}}^* + o_p(n^{-1/2})$$

and the variance of $\tilde{\theta}_{\text{opt}}^*$ is equal to V_l , the lower bound of the asymptotic variance.

The (asymptotic) optimality of the estimator in (17) is justified under the joint distribution of the response model (4) and the superpopulation model (16). When both models are correct, $\hat{\theta}_{\text{opt}}$ is optimal and the choice of $(\hat{\beta}, \hat{\phi})$ does not affect the efficiency of the $\hat{\theta}_{\text{opt}}$ as long as $(\hat{\beta}, \hat{\phi})$ is \sqrt{n} -consistent. Robins, Rotnitzky and Zhao (1994) also advocated using $\hat{\theta}_{\text{opt}}$ in (17) under simple random sampling.

Remark 1 When the response model is correct and the superpopulation model (16) is not necessarily correct, the choice of $\hat{\beta}$ does affect the efficiency of the optimal estimator. Cao, Tsiatis and Davidian (2009) considered optimal estimation when only the response model is correct. Using Taylor linearization, the optimal estimator in (17) with $\hat{\phi}$ satisfying (6) is asymptotically equivalent to

$$\tilde{\theta}(\beta) = \sum_{i \in A} w_i \left[m(\mathbf{x}_i; \beta) + \frac{r_i}{p_i} \{y_i - m(\mathbf{x}_i; \beta)\} - \left(\frac{r_i}{p_i} - 1 \right) \mathbf{c}_{\beta}' p_i \mathbf{h}_i \right],$$

where \mathbf{c}_{β} is the probability limit of $\hat{\mathbf{c}}_{\beta} = \{\sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) \hat{p}_i \mathbf{h}_i'(\hat{\phi})\}^{-1} \sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) \{y_i - m(\mathbf{x}_i; \beta)\}$ and $\mathbf{z}_i(\phi) = \partial \{p_i^{-1}(\phi)\} / \partial \phi$. The asymptotic variance is then equal to

$$V\{\tilde{\theta}(\beta)\} =$$

$$V(\hat{\theta}_{\text{HT}}) + E \left[\sum_{i \in A} w_i^2 \frac{1 - \hat{p}_i}{\hat{p}_i} \{y_i - m(\mathbf{x}_i; \beta) - \mathbf{c}_{\beta}' p_i \mathbf{h}_i\}^2 \right].$$

Thus, an optimal estimator of β can be computed by finding $\hat{\beta}$ that minimizes

$$Q(\beta) = \sum_{i \in A} w_i^2 r_i \frac{1 - \hat{p}_i}{\hat{p}_i^2} \{y_i - m(\mathbf{x}_i; \beta) - \hat{\mathbf{c}}_{\beta}' \hat{p}_i \mathbf{h}_i(\hat{\phi})\}^2.$$

The resulting estimator is design-optimal in the sense that it minimizes the asymptotic variance under the response model.

3. Augmented propensity score model

In this section, we consider optimal PSA estimation. Note that the optimal estimator $\hat{\theta}_{\text{opt}}$ in (17) is not necessarily written as a PSA estimator form in (3). It is in the PSA estimator form if it satisfies $\sum_{i \in A} w_i r_i \hat{p}_i^{-1} m(\mathbf{x}_i; \hat{\beta}) = \sum_{i \in A} w_i m(\mathbf{x}_i; \hat{\beta})$. Thus, we can construct an optimal PSA estimator by including $m(\mathbf{x}_i; \hat{\beta})$ in the model for the propensity score. Specifically, given $\hat{m}_i = m(\mathbf{x}_i; \hat{\beta})$, $\hat{p}_i = p_i(\hat{\phi})$ and $\hat{\mathbf{h}}_i = \mathbf{h}_i(\hat{\phi})$, where $\hat{\phi}$ is obtained from (6), we augment the response model by

$$p_i^*(\hat{\phi}, \lambda) = \frac{\hat{p}_i}{\hat{p}_i + (1 - \hat{p}_i) \exp(\lambda_0 + \lambda_1 \hat{m}_i)}, \quad (18)$$

where $\lambda = (\lambda_0, \lambda_1)'$ is the Lagrange multiplier which is used to incorporate the additional constraint. If $(\lambda_0, \lambda_1)' = \mathbf{0}$, then $p_i^*(\hat{\phi}, \lambda) = \hat{p}_i$. The augmented response probability $p_i^*(\hat{\phi}, \lambda)$ always takes values between 0 and 1. The augmented response probability model (18) can be derived by minimizing the Kullback-Leibler distance $\sum_{i \in A} w_i r_i q_i^* \log(q_i^*/q_i)$, where $q_i^* = (1 - p_i^*)/p_i^*$ and $q_i = (1 - \hat{p}_i)/\hat{p}_i$, subject to the constraint $\sum_{i \in A} w_i (r_i/p_i^*)(1, \hat{m}_i) = \sum_{i \in A} w_i (1, \hat{m}_i)$.

Using (18), the optimal PSA estimator is computed by

$$\hat{\theta}_{\text{PSA}}^* = \frac{1}{N} \sum_{i \in A} w_i \frac{r_i}{p_i^*(\hat{\phi}, \hat{\lambda})} y_i, \quad (19)$$

where $\hat{\lambda}$ satisfies

$$\sum_{i \in A} w_i \frac{r_i}{p_i^*(\hat{\phi}, \hat{\lambda})} (1, \hat{m}_i) = \sum_{i \in A} w_i (1, \hat{m}_i). \quad (20)$$

Under the response model (4), it can be shown that

$$\hat{\theta}_{\text{PSA}}^* = \frac{1}{N} \sum_{i \in A} w_i \left\{ \hat{b}_0 + \hat{b}_1 \hat{m}_i + \frac{r_i}{\hat{p}_i} (y_i - \hat{b}_0 - \hat{b}_1 \hat{m}_i) \right\} + o_p(n^{-1/2}),$$

where

$$\begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix} = \left\{ \sum_{i \in A} w_i r_i \left(\frac{1}{\hat{p}_i} - 1 \right) \begin{pmatrix} 1 \\ \hat{m}_i \end{pmatrix} \begin{pmatrix} 1 \\ \hat{m}_i \end{pmatrix}' \right\}^{-1} \sum_{i \in A} w_i r_i \left(\frac{1}{\hat{p}_i} - 1 \right) \begin{pmatrix} 1 \\ \hat{m}_i \end{pmatrix} y_i. \quad (21)$$

Furthermore, by the argument for Theorem 1, we can establish that

$$\begin{aligned} \hat{\theta}_{\text{PSA}}^* &= \frac{1}{N} \sum_{i \in A} w_i \left\{ b_0 + b_1 \hat{m}_i + \gamma'_{h2} p_i \mathbf{h}_i \right. \\ &\quad \left. + \frac{r_i}{p_i} (y_i - b_0 - b_1 \hat{m}_i - \gamma'_{h2} p_i \mathbf{h}_i) \right\} \\ &\quad + o_p(n^{-1/2}), \end{aligned}$$

where (b_0, b_1, γ'_{h2}) is the probability limit of $(\hat{b}_0, \hat{b}_1, \hat{\gamma}'_{h2})$ with

$$\begin{aligned} \hat{\gamma}_{h2} &= \left\{ \sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) \hat{p}_i \mathbf{h}_i'(\hat{\phi}) \right\}^{-1} \\ &\quad \sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) (y_i - \hat{b}_0 - \hat{b}_1 \hat{m}_i) \end{aligned} \quad (22)$$

and the effect of estimating ϕ_0 in $\hat{p}_i = p(\mathbf{x}_i; \hat{\phi})$ can be safely ignored.

Note that, under the response model (4), $(\hat{\phi}, \hat{\lambda})$ in (19) converges in probability to $(\phi_0, \mathbf{0})$, where ϕ_0 is the true parameter in (4). Thus, the propensity score from the augmented model converges to the true response probability.

Because $\hat{\lambda}$ converges to zero in probability, the choice of $\hat{\beta}$ in $\hat{m}_i = m(\mathbf{x}_i; \hat{\beta})$ does not play a role for the asymptotic unbiasedness of the PSA estimator. The asymptotic variances are changed for different choices of $\hat{\beta}$.

Under the superpopulation model (16), $\hat{b}_0 + \hat{b}_1 \hat{m}_i \rightarrow E(Y | \mathbf{x}_i)$ in probability. Thus, the optimal PSA estimator in (19) is asymptotically equivalent to the optimal estimator in (17). Incorporating \hat{m}_i into the calibration equation to achieve optimality is close in spirit to the model-calibration method proposed by Wu and Sitter (2001).

4. Variance estimation

We now discuss variance estimation of PSA estimators under the assumed response model. Singh and Folsom (2000) and Kott (2006) discussed variance estimation for certain types of PSA estimators. Kim and Kim (2007) discussed variance estimation when the PSA estimator is computed with the maximum likelihood method.

We consider variance estimation for the PSA estimator of the form (3) where $\hat{p}_i = p_i(\hat{\phi})$ is constructed to satisfy (6) for some $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi, \beta)$, where β^* is obtained using the postulated superpopulation model. Let β^* be the probability limit of $\hat{\beta}$ under the response model. Note that β^* is not necessarily equal to β_0 in (16) since we are not assuming that the postulated superpopulation model is correctly specified in this section.

Using the argument for the Taylor linearization (13) used in the proof of Theorem 1, the PSA estimator satisfies

$$\hat{\theta}_{\text{PSA}} = \frac{1}{N} \sum_{i \in A} w_i \eta_i(\phi_0, \beta^*) + o_p(n^{-1/2}), \quad (23)$$

where

$$\begin{aligned} \eta_i(\phi, \beta) &= p_i(\phi) \mathbf{h}_i'(\phi, \beta) \gamma_h^* \\ &\quad + \frac{r_i}{p_i(\phi)} \{ y_i - p_i(\phi) \mathbf{h}_i'(\phi, \beta) \gamma_h^* \}, \end{aligned} \quad (24)$$

$\mathbf{h}_i(\phi, \beta) = \mathbf{h}(\mathbf{x}_i; \phi, \beta)$ and γ_h^* is defined as in (8) with \mathbf{h}_i replaced by $\mathbf{h}_i(\phi_0, \beta^*)$. Since $p_i(\hat{\phi})$ satisfies (6) with $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi, \beta)$, $\hat{\theta}_{\text{PSA}} = N^{-1} \sum_{i \in A} w_i \eta_i(\hat{\phi}, \hat{\beta})$ holds and the linearization in (23) can be expressed as $N^{-1} \sum_{i \in A} w_i \eta_i(\hat{\phi}, \hat{\beta}) = N^{-1} \sum_{i \in A} w_i \eta_i(\phi_0, \beta^*) + o_p(n^{-1/2})$. Thus, if (\mathbf{x}_i, y_i, r_i) are independent and identically distributed (IID), then $\eta_i(\phi_0, \beta^*)$ are IID even though $\eta_i(\hat{\phi}, \hat{\beta})$ are not necessarily IID. Because $\eta_i(\phi_0, \beta^*)$ are IID, we can apply the standard complete sample method to estimate the variance of $\hat{\eta}_{\text{HT}} = N^{-1} \sum_{i \in A} w_i \eta_i(\phi_0, \beta^*)$, which is asymptotically equivalent to the variance of $\hat{\theta}_{\text{PSA}} = N^{-1} \sum_{i \in A} w_i \eta_i(\hat{\phi}, \hat{\beta})$. See Kim and Rao (2009).

To derive the variance estimator, we assume that the variance estimator $\hat{V} = N^{-2} \sum_{i \in A} \sum_{j \in A} \Omega_{ij} g_i g_j$ satisfies

$\hat{V}/V(\hat{g}_{HT}|\mathcal{F}_N) = 1 + o_p(1)$ for some Ω_{ij} related to the joint inclusion probability, where $\hat{g}_{HT} = N^{-1}\sum_{i \in A} w_i g_i$ for any g with a finite second moment and $V(g_{HT}|\mathcal{F}_N) = N^{-2}\sum_{i=1}^N \sum_{j=1}^N \Omega_{N-ij} g_i g_j$, for some Ω_{N-ij} . We also assume that

$$\sum_{i=1}^N |\Omega_{N-ij}| = O(n^{-1}N). \quad (25)$$

To obtain the total variance, the *reverse framework* of Fay (1992), Shao and Steel (1999), and Kim and Rao (2009) is considered. In this framework, the finite population is first divided into two groups, a population of respondents and a population of nonrespondents. Given the population, the sample A is selected according to a probability sampling design. Thus, selection of the population respondents from the whole finite population is treated as the first-phase sampling and the selection of the sample respondents from the population respondents is treated as the second-phase sampling in the reverse framework. The total variance of $\hat{\eta}_{HT}$ can be written as

$$V(\hat{\eta}_{HT}|\mathcal{F}_N) = V_1 + V_2 = E\{V(\hat{\eta}_{HT}|\mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\} + V\{E(\hat{\eta}_{HT} | \mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\}. \quad (26)$$

The conditional variance term $V(\hat{\eta}_{HT} | \mathcal{F}_N, \mathcal{R}_N)$ in (26) can be estimated by

$$\hat{V}_1 = N^{-2} \sum_{i \in A} \sum_{j \in A} \Omega_{ij} \hat{\eta}_i \hat{\eta}_j, \quad (27)$$

where $\hat{\eta}_i = \eta_i(\hat{\phi}, \hat{\beta})$ is defined in (24) with γ_h^* replaced by a consistent estimator such as $\hat{\gamma}_h^* = \{\sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) \hat{p}_i \hat{\mathbf{h}}_i'\}^{-1} \sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) y_i$, and $\hat{\mathbf{h}}_i = \mathbf{h}(\mathbf{x}_i; \hat{\phi}, \hat{\beta})$. To show that \hat{V}_1 is also consistent for V_1 in (26), it suffices to show that $V\{n \cdot V(\hat{\eta}_{HT} | \mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\} = o(1)$, which follows by (25) and the existence of the fourth moment. See Kim, Navarro and Fuller (2006). The second term V_2 in (26) is

$$V\{E(\hat{\eta}_{HT} | \mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\} = V\left(N^{-1} \sum_{i=1}^N \eta_i | \mathcal{F}_N\right) = \frac{1}{N^2} \sum_{i=1}^N \frac{1-p_i}{p_i} (y_i - p_i \mathbf{h}_i^* \gamma_h^*)^2,$$

where $\mathbf{h}_i^* = \mathbf{h}(\mathbf{x}_i; \phi_0, \beta^*)$. A consistent estimator of V_2 can be derived as

$$\hat{V}_2 = \frac{1}{N^2} \sum_{i \in A} w_i r_i \frac{1-\hat{p}_i}{\hat{p}_i^2} (y_i - \hat{p}_i \hat{\mathbf{h}}_i' \hat{\gamma}_h^*)^2, \quad (28)$$

where $\hat{\gamma}_h^*$ is defined after (27). Therefore,

$$\hat{V}(\hat{\theta}_{PSA}) = \hat{V}_1 + \hat{V}_2, \quad (29)$$

is consistent for the variance of the PSA estimator defined in (3) with $\hat{p}_i = p_i(\hat{\phi})$ satisfying (6), where \hat{V}_1 is in (27) and \hat{V}_2 is in (28).

Note that the first term of the total variance is $V_1 = O_p(n^{-1})$, but the second term is $V_2 = O_p(N^{-1})$. Thus, when the sampling fraction nN^{-1} is negligible, that is, $nN^{-1} = o(1)$, the second term V_2 can be ignored and \hat{V}_1 is a consistent estimator of the total variance. Otherwise, the second term V_2 should be taken into consideration, so that a consistent variance estimator can be constructed as in (29).

Remark 2 The variance estimation of the optimal PSA estimator with augmented propensity model (18) with $(\hat{\phi}, \hat{\lambda})$ satisfying (20) can be derived by (29) using $\hat{\eta}_i = \hat{b}_0 + \hat{b}_1 \hat{m}_i + \hat{\gamma}_{h2} \hat{p}_i \hat{\mathbf{h}}_i + r_i \hat{p}_i^{-1} (y_i - \hat{b}_0 - \hat{b}_1 \hat{m}_i - \hat{\gamma}_{h2} \hat{p}_i \hat{\mathbf{h}}_i)$ where (\hat{b}_0, \hat{b}_1) and $\hat{\gamma}_{h2}$ are defined in (21) and (22), respectively.

5. Simulation study

5.1 Study one

Two simulation studies were performed to investigate the properties of the proposed method. In the first simulation, we generated a finite population of size $N = 10,000$ from the following multivariate normal distribution:

$$\begin{pmatrix} x_1 \\ x_2 \\ e \end{pmatrix} \sim N \left[\begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right].$$

The variable of interest y was constructed as $y = 1 + x_1 + e$. We also generated response indicator variables r_i independently from a Bernoulli distribution with probability

$$p_i = \frac{\exp(2 + x_{2i})}{1 + \exp(2 + x_{2i})}.$$

From the finite population, we used simple random sampling to select two samples of size, $n = 100$ and $n = 400$, respectively. We used $B = 5,000$ Monte Carlo samples in the simulation. The average response rate was about 69.6%.

To compute the propensity score, a response model of the form

$$p(\mathbf{x}; \phi) = \frac{\exp(\phi_0 + \phi_1 x_2)}{1 + \exp(\phi_0 + \phi_1 x_2)} \quad (30)$$

was postulated and an outcome regression model of the form

$$m(\mathbf{x}; \beta) = \beta_0 + \beta_1 x_1 \quad (31)$$

was postulated to obtain the optimal PSA estimators. Thus, both models are correctly specified. From each sample, we computed four estimators of $\theta = N^{-1} \sum_{i=1}^N y_i$:

- 1. (PSA-MLE): PSA estimator in (3) with $\hat{p}_i = p_i(\hat{\phi})$ and $\hat{\phi}$ being the maximum likelihood estimator of ϕ .
- 2. (PSA-CAL): PSA estimator in (3) with \hat{p}_i satisfying the calibration constraint (15) on $(1, x_{2i})$.
- 3. (AUG): Augmented PSA estimator in (19).
- 4. (OPT): Optimal estimator in (17).

In the augmented PSA estimators, $\hat{\phi}$ was computed by the maximum likelihood method. Under model (30), the maximum likelihood estimator of $\phi = (\phi_0, \phi_1)'$ was computed by solving (6) with $\mathbf{h}_i(\phi) = (1, x_{2i})'$. Parameter (β_0, β_1) for the outcome regression model was computed using ordinary least squares, regressing y on x_1 . In addition to the point estimators, we also computed the variance estimators of the point estimators. The variance estimators of the PSA estimators were computed using the pseudo-values in (24) and the $\mathbf{h}_i(\phi)$ corresponding to each estimator. For the augmented PSA estimators, the pseudo-values were computed by the method in Remark 2.

Table 1 presents the Monte Carlo biases, variances, and mean square errors of the four point estimators and the Monte Carlo percent relative biases and t -statistics of the variance estimators of the estimators. The percent relative bias of a variance estimator $\hat{V}(\hat{\theta})$ is calculated as $100 \times \{V_{MC}(\hat{\theta})\}^{-1} [E_{MC}\{\hat{V}(\hat{\theta})\} - V_{MC}(\hat{\theta})]$, where $E_{MC}(\cdot)$ and $V_{MC}(\cdot)$ denote the Monte Carlo expectation and the Monte Carlo variance, respectively. The t -statistic in Table 1 is the test statistic for testing the zero bias of the variance estimator. See Kim (2004).

Based on the simulation results in Table 1, we have the following conclusions.

- 1. All of the PSA estimators are asymptotically unbiased because the response model (30) is correctly specified. The PSA estimator using the calibration method is slightly more efficient than the PSA estimator using the maximum likelihood estimator, because the last term of (14) is smaller for the calibration method as the predictor for $E(Y | \mathbf{x}_i) = \beta_0 + \beta_1 x_{1i}$ is better approximated by a linear function of $(1, x_{2i})$ than by a linear function of $(\hat{p}_i, \hat{p}_i x_{2i})$.

- 2. The augmented PSA estimator is more efficient than the direct PSA estimator (3). The augmented PSA estimator is constructed by using the correctly specified regression model (31) and so it is asymptotically equivalent to the optimal PSA estimator in (17).
- 3. Variance estimators are all approximately unbiased. There are some modest biases in the variance estimators of the PSA estimators when the sample size is small ($n = 100$).

5.2 Study two

In the second simulation study, we further investigated the PSA estimators with a non-linear outcome regression model under an unequal probability sampling design. We generated two stratified finite populations of (x, y) with four strata ($h = 1, 2, 3, 4$), where x_{hi} were independently generated from a normal distribution $N(1, 1)$ and y_{hi} were dichotomous variables that take values of 1 or 0 from a Bernoulli distribution with probability p_{1yhi} or p_{2yhi} . Two different probabilities were used for two populations, respectively:

- 1. Population 1 (Pop1):

$$p_{1yhi} = 1 / \{1 + \exp(0.5 - 2x)\}.$$

- 2. Population 2 (Pop2):

$$p_{2yhi} = 1 / [1 + \exp\{0.25(x - 1.5)^2 - 1.5\}].$$

In addition to x_{hi} and y_{hi} , the response indicator variables r_{hi} were generated from a Bernoulli distribution with probability $p_{hi} = 1 / \{1 + \exp(-1.5 + 0.7x_{hi})\}$. The sizes of the four strata were $N_1 = 1,000$, $N_2 = 2,000$, $N_3 = 3,000$, and $N_4 = 4,000$, respectively. In each of the two sets of finite population, a stratified sample of size $n = 400$ was independently generated without replacement, where a simple random sample of size $n_h = 100$ was selected from each stratum. We used $B = 5,000$ Monte Carlo samples in this simulation. The average response rate was about 67%.

Table 1
Monte Carlo bias, variance and mean square error(MSE) of the four point estimators and percent relative biases (R.B.) and t -statistics(t -stat) of the variance estimators based on 5,000 Monte Carlo samples

n	Method	$\hat{\theta}$			$V(\hat{\theta})$	
		Bias	Variance	MSE	R.B. (%)	t-stat
100	(PSA-MLE)	-0.01	0.0315	0.0317	-2.34	-1.12
	(PSA-CAL)	-0.01	0.0308	0.0309	-3.56	-1.70
	(AUG)	0.00	0.0252	0.0252	-0.61	-0.30
	(OPT)	0.00	0.0252	0.0252	-0.21	-0.10
400	(PSA-MLE)	-0.01	0.00737	0.00746	0.35	0.17
	(PSA-CAL)	-0.01	0.00724	0.00728	0.29	0.14
	(AUG)	0.00	0.00612	0.00612	0.07	0.03
	(OPT)	0.00	0.00612	0.00612	-0.14	-0.07

To compute the propensity score, a response model of the form

$$p(x; \phi) = \frac{\exp(\phi_0 + \phi_1 x)}{1 + \exp(\phi_0 + \phi_1 x)}$$

was postulated for parameter estimation. To obtain the augmented PSA estimator, a model for the variable of interest of the form

$$m(x; \beta) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{32}$$

was postulated. Thus, model (32) is a true model under (Pop1), but it is not a true model under (Pop2).

We computed four estimators:

1. (PSA-MLE): PSA estimator in (3) using the maximum likelihood estimator of ϕ .
2. (PSA-CAL): PSA estimator in (3) with \hat{p}_i satisfying the calibration constraint (15) on $(1, x)$.
3. (AUG-1): Augmented PSA estimator $\hat{\theta}_{\text{PSA}}^*$ in (19) with $\hat{\beta}$ computed by the maximum likelihood method.
4. (AUG-2): Augmented PSA estimator $\hat{\theta}_{\text{PSA}}^*$ in (19) with $\hat{\beta}$ computed by the method of Cao *et al.* (2009) discussed in Remark 1.

We considered the the augmented PSA estimator in (19) with $\hat{p}_i = p_i(\hat{\phi})$, where $\hat{\phi}$ is the maximum likelihood estimator of ϕ . The first augmented PSA estimator (AUG-1) used $\hat{m}_i = m(x_i; \hat{\beta})$ with $\hat{\beta}$ found by solving $\sum_{h=1}^4 \sum_{i \in A_h} w_{hi} r_{hi} \{y_{hi} - m(x_{hi}; \beta)\}(1, x_{hi}) = \mathbf{0}$, where A_h is the set of indices appearing in the sample for stratum h and w_{hi} is the sampling weight of unit i for stratum h .

Table 2 presents the simulation results for each method. In each population, the augmented PSA estimator shows some improvement comparing to the PSA estimator using the maximum likelihood estimator of ϕ or the calibration estimator of ϕ in terms of variance. Under (Pop1), since model (32) is true, there is essentially no difference between

the augmented PSA estimators using different methods of estimating β . However, under (Pop2), where the assumed outcome regression model (32) is incorrect, the augmented PSA estimator with $\hat{\beta}$ computed by the method of Cao *et al.* (2009) results in slightly better efficiency, which is consistent with the theory in Remark 1. Variance estimates are approximately unbiased in all cases in the simulation study.

6. Conclusion

We have considered the problem of estimating the finite population mean of y under nonresponse using the propensity score method. The propensity score is computed from a parametric model for the response probability, and some asymptotic properties of PSA estimators are discussed. In particular, the optimal PSA estimator is derived with an additional assumption for the distribution of y . The propensity score for the optimal PSA estimator can be implemented by the augmented propensity model presented in Section 3. The resulting estimator is still consistent even when the assumed outcome regression model fails to hold.

We have restricted our attention to missing-at-random mechanisms in which the response probability depends only on the always-observed \mathbf{x} . If the response mechanism also depends on y , PSA estimation becomes more challenging. PSA estimation when missingness is not at random is beyond the scope of this article and will be a topic of future research.

Acknowledgements

The research was partially supported by a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University. The authors wish to thank F. Jay Breidt, three anonymous referees, and the associate editor for their helpful comments.

Table 2
Monte Carlo bias, variance and mean square error of the four point estimators and percent relative biases (R.B.) and *t*-statistics of the variance estimators, based on 5,000 Monte Carlo samples

Population	Method	$\hat{\theta}_{\text{PSA}}$			$V(\hat{\theta}_{\text{PSA}})$	
		Bias	Variance	MSE	R.B. (%)	<i>t</i> -stat
Pop1	(PSA-MLE)	0.00	0.000750	0.000762	-1.13	-0.57
	(PSA-CAL)	0.00	0.000762	0.000769	-1.45	-0.72
	(AUG-1)	0.00	0.000745	0.000757	-1.73	-0.86
	(AUG-2)	0.00	0.000745	0.000757	-1.83	-0.91
Pop2	(PSA-MLE)	0.00	0.000824	0.000826	0.29	0.14
	(PSA-CAL)	0.00	0.000829	0.000835	-0.94	-0.46
	(AUG-1)	0.00	0.000822	0.000823	-0.71	-0.35
	(AUG-2)	0.00	0.000820	0.000821	-0.61	-0.30

References

- Cao, W., Tsatis, A.A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96, 723-734.
- Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34, 305-334.
- Da Silva, D.N., and Opsomer, J.D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *Canadian Journal of Statistics*, 34, 563-579.
- Da Silva, D.N., and Opsomer, J.D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, 35, 2, 165-176.
- Duncan, K.B., and Stasny, E.A. (2001). Using propensity scores to control coverage bias in telephone surveys. *Survey Methodology*, 27, 2, 121-130.
- Durrant, G.B., and Skinner, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Survey Methodology*, 32, 1, 25-36.
- Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.
- Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistics Section*, American Statistical Association, 197-202.
- Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 1, 75-85.
- Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 637-642.
- Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- Kim, J.K. (2004). Finite sample properties of multiple imputation estimators. *The Annals of Statistics*, 32, 766-783.
- Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35, 501-514.
- Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Kim, J.K., and Rao, J.N.K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.
- Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 2, 133-142.
- Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22, 329-349.
- Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-296.
- Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business and Economic Statistics*, 21, 43-52.
- Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.
- Randles, R.H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, 10, 462-474.
- Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 1, 43-53.
- Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.
- Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387-394.
- Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Singh, A.C., and Folsom, R.E. (2000). Bias corrected estimating function approach for variance estimation adjusted for poststratification. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 610-615.
- Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Assessing the accuracy of response propensity models in longitudinal studies

Ian Plewis, Sosthenes Ketende and Lisa Calderwood¹

Abstract

Non-response in longitudinal studies is addressed by assessing the accuracy of response propensity models constructed to discriminate between and predict different types of non-response. Particular attention is paid to summary measures derived from receiver operating characteristic (ROC) curves and logit rank plots. The ideas are applied to data from the UK Millennium Cohort Study. The results suggest that the ability to discriminate between and predict non-respondents is not high. Weights generated from the response propensity models lead to only small adjustments in employment transitions. Conclusions are drawn in terms of the potential of interventions to prevent non-response.

Key Words: Longitudinal studies; Missing data; Weighting; Propensity scores; ROC curves; Millennium Cohort Study.

1. Introduction

Examples of studies that have modelled the predictors of different kinds of, and different reasons for the non-response that affect longitudinal studies are plentiful, stimulated by being able to draw on auxiliary variables obtained from sample members before (and after) the occasions at which they are non-respondents. See, for example, Lepkowski and Couper (2002) for an analysis that separates refusals from not being located or contacted; Hawkes and Plewis (2006) who separate wave non-respondents from attrition cases in the UK National Child Development Study; and Plewis (2007a) and Plewis, Ketende, Joshi and Hughes (2008) who consider non-response in the first two waves of the UK Millennium Cohort Study. The focus of this paper is on how we can assess the accuracy of these response propensity models (Little and Rubin 2002). The paper is built around a framework that is widely used in epidemiology (Pepe 2003) and criminology (Copas 1999) to evaluate risk scores but has not, to our knowledge, been used in survey research before. Response propensity models can be used to construct weights intended to remove biases from estimates, to inform imputations, and to predict potential non-respondents at future waves thereby directing fieldwork resources to those respondents who might otherwise be lost. The accuracy of response propensity models has not, however, been given the amount of attention it warrants in terms of their ability to discriminate between respondents and non-respondents, and to predict future non-response. Good estimates of accuracy can be used to compare the efficacy of different weighting methods, and to help to determine the allocation of scarce fieldwork resources in order to reduce non-response.

The paper is organised as follows. The framework for assessing accuracy is set out in the next section. Section 3 introduces the UK Millennium Cohort Study and the methods are illustrated using data from this study in Section 4. Section 5 concludes.

2. Models for predicting non-response

A typical response propensity model for a binary outcome (*e.g.*, Hawkes and Plewis 2006) is:

$$f(\pi_{it}) = \sum_p \beta_p x_{pi} + \sum_q \gamma_q x_{qi,t-k}^* + \sum_r \sum_k \delta_r z_{ri,t-k} \quad (1)$$

where

- $\pi_{it} = E(r_{it})$ is the probability of not responding for subject i at wave t ; $r_{it} = 0$ for a response and 1 for non-response; f is an appropriate function such as logit or probit.
- $i = 1, \dots, n$ where n is the observed sample size at wave one.
- $t = 1, \dots, T_i$ where T_i is the number of waves for which r_{it} is recorded for subject i .
- x_{pi} are fixed characteristics of subject i measured at wave one, $p = 0, \dots, P$; $x_0 = 1$ for all i .
- $x_{qi,t-k}^*$ are time-varying characteristics of subject i , measured at waves $t - k$, $q = 1, \dots, Q$, $k = 1, 2, \dots$, often k will be 1.
- $z_{ri,t-k}$ are time-varying characteristics of the data collection process, measured for subject i at waves $t - k$, $r = 1, \dots, R$, $k = 0, 1, \dots$, often k will be 1 but can be 0 for variables such as number of contacts before a response is obtained.

Model (1) can easily be extended to more than two response categories such as {response, wave non-response, attrition}. Other approaches are also possible. For example, it is often more convenient to model the probability of not responding just at wave $t = t^*$ in terms of variables measured at earlier waves $t^* - k$, $k \geq 1$ or, when there is no wave non-response so that non-response has a monotonic rather than an arbitrary pattern, to model time to attrition as a survival process.

The estimated response probabilities p_t , for $t = t^*$, are derived from the estimated non-response probabilities in (1) and they can be used to generate inverse probability weights $g_t (= 1/p_t)$. These are widely applied (see Section 4.2 for an example) to adjust for biases arising from non-response under the assumption that data are missing at random (MAR) as defined by Little and Rubin (2002).

2.1 Assessing the accuracy of predictions

A widely used method of assessing the accuracy of models like (1) is to estimate their goodness-of-fit by using one of several possible pseudo- R^2 statistics. Estimates of pseudo- R^2 are not especially useful in this context, partly because they are difficult to compare across datasets but also because they assess the overall fit of the model and do not, therefore, distinguish between the accuracy of the model for the respondents and non-respondents separately.

As Pepe (2003) emphasises, there are two related components of accuracy: discrimination (or classification) and prediction. Discrimination refers to the conditional probabilities of having a propensity score (s : the linear predictor from (1)) above a chosen threshold (c) given that a person either is or is not a non-respondent. Prediction, on the other hand, refers to the conditional probabilities of being or

becoming a non-respondent given a propensity score above or below the threshold.

More formally, let D and \bar{D} refer to the presence and absence of the poor outcome (*i.e.*, non-response) and define $+$ ($s > c$) and $-$ ($s \leq c$) as positive and negative tests derived from the propensity score and its threshold. Then, for discrimination, we are interested in $P(+|D)$, the true positive fraction (TPF) or sensitivity of the test, and $P(-|\bar{D})$ its specificity, equal to one minus the false positive fraction ($1 - \text{FPF}$). For prediction, however, we are interested in $P(D|+)$, the positive predictive value (PPV) and $P(\bar{D}|-)$, the negative predictive value (NPV). If the probability of a positive test ($P(+)$) is the same as the prevalence of the poor outcome ($P(D) = \rho$) then inferences about discrimination and prediction are essentially the same: sensitivity equals PPV and specificity equals NPV. Generally, however, $\{\text{TPF}, \text{FPF}, \rho\}$ and $\{\text{PPV}, \text{NPV}, \tau\}$ convey different pieces of information. TPF can be plotted against FPF for any risk score threshold c . This is the receiver operating characteristic (ROC) curve (Figure 1). Krzanowski and Hand (2009) give a detailed discussion of how to estimate ROC curves. The AUC – the area enclosed by the ROC curve and the x-axis in Figure 1 – is of particular interest and can vary from 1 (perfect discrimination) down to 0.5, the area below the diagonal (implying no discrimination). The AUC can be interpreted as the probability of assigning a pair of cases, one respondent and one non-respondent, to their correct categories, bearing in mind that guessing would correspond to a probability of 0.5. A linear transformation of AUC ($= 2 \cdot \text{AUC} - 1$) – sometimes referred to as a Gini coefficient and equivalent to Somer's D rank correlation index (Harrell, Lee and Mark 1996) – is commonly used as a more natural measure than AUC because it varies from 0 to 1.

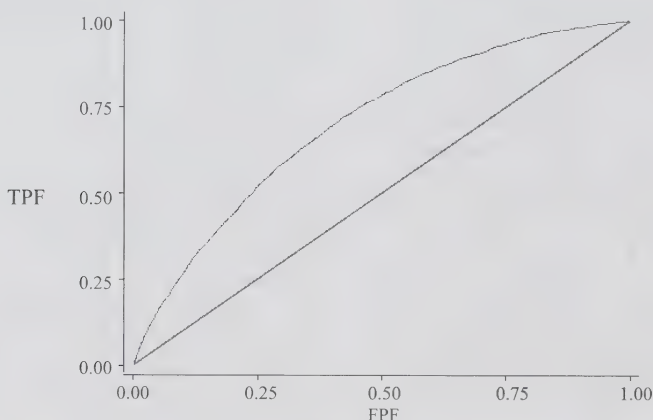


Figure 1 ROC curve

Copas (1999) proposes the logit rank plot as an alternative to the ROC as a means of assessing the predictiveness of a propensity score. If the propensity score is derived from a logistic regression then a logit rank plot is just a plot of the linear predictor from the model against the logistic transformation of the proportional rank of the propensity scores. More generally, it is a plot of $\text{logit}(p_i)$ where p_i is the estimated probability from any form of (1) i.e., $p(D|\mathbf{x}, \mathbf{x}^*, \mathbf{z})$, against the logits of the proportional ranks (r/n) where r is the rank position of case i ($i = 1, \dots, n$) on the propensity score. This relation is usually close to being linear and its slope – which can vary from zero to one – is a measure of the predictive strength of the propensity score. Copas argues that the slope is more sensitive to changes in the specification of the propensity model, and to changes in the prevalence of the outcome, than the Gini coefficient is. A good estimate of the slope can be obtained by calculating quantiles of the variables on the y and x axes and then fitting a simple regression model.

The extent to which propensity scores discriminate between respondents and non-respondents is one indicator of the effectiveness of any statistical adjustments for missingness. A lack of discrimination suggests either that there are important predictors absent from the propensity score or that a substantial part of the process that drives the missingness is essentially random. The extent to which propensity scores predict whether a case will be a non-respondent in subsequent waves – and what kind of non-respondent they will be – is an indication of whether any intervention to reduce non-response will be successful.

3. The Millennium Cohort Study

The wave one sample of the UK Millennium Cohort Study (MCS) includes 18,552 families born over a 12-month period during the years 2000 and 2001, and living in selected UK electoral wards at age nine months. The initial response rate was 72%. Areas with high proportions of

Black and Asian families, disadvantaged areas and the three smaller UK countries are all over-represented in the sample which is disproportionately stratified and clustered as described in Plewis (2007b). The first four waves took place when the cohort members were (approximately) nine months, 3, 5 and 7 years old. At wave two, 19% of the target sample – which excludes child deaths and emigrants – were unproductive. The unproductive cases were equally divided between wave non-response and attrition, and between refusals and other non-productives (not located, not contacted etc.).

4. Analyses of non-response

4.1 Accuracy of discrimination and prediction

Plewis (2007a) and Plewis *et al.* (2008) show that variables measured at wave one of the MCS that are associated with attrition at wave two are not necessarily associated with wave non-response then (and vice-versa). The same is true for correlates of refusal and other non-productives. Table 1 gives the accuracy estimates from the response propensity models. The estimate of the Gini coefficient for overall non-response (0.38) is relatively low: it corresponds to an AUC of 0.69 which is the probability of correctly assigning (based on their predicted probabilities) a pair of cases (one respondent, one non-respondent), indicating that discrimination between non-respondents and respondents from the propensity score is not especially good. Discrimination is slightly better for wave non-respondents than it is for attrition and notably better for other non-productive than it is for refusal. These estimates were obtained from pairwise comparisons of each non-response category with being a respondent. A similar picture emerges when we look at the slopes of the logit rank plots although these bring out more clearly the differences in predictiveness for the different types of, and reasons for non-response.

Table 1
Accuracy estimates from response propensity models, MCS wave two

Accuracy measure	Overall non-response ⁽²⁾	Non-response type ⁽²⁾		Non-response reason ⁽²⁾	
		Wave non-response	Attrition	Refusal	Other non-productive
AUC ⁽¹⁾	0.69	0.71	0.69	0.68	0.77
Gini ⁽¹⁾	0.38	0.42	0.39	0.37	0.53
Logit rank plot: slope ⁽¹⁾	0.45	0.51	0.44	0.40	0.63
Sample size	18,230	16,210	16,821	16,543	16,513

⁽¹⁾ AUC estimated under the binormal assumption (Krzanowski and Hand 2009); 95% confidence limits for (a) AUC not more than ± 0.015 , (b) Gini coefficient and logit rank plot slope not more than ± 0.03 .
⁽²⁾ Based on a logistic regression, allowing for the survey design using the *svy* commands in STATA with the sample size based on the sum of the productive and relevant non-response category.

The correct specification of models for explaining non-response can be difficult to achieve. New candidates for inclusion in a model can appear after the model and the corresponding inverse probability weights have been estimated, others remain unknown. How much effect on measures of accuracy might the inclusion of new variables have? Here we examine the effects of adding three new variables to the MCS models: (i) whether or not respondents gave consent to having their survey records linked to health records at wave one; (ii) a neighbourhood conditions score derived from interviewer observations at wave two; and (iii) whether, at wave one, the main respondent reported voting at the last UK general election. The first two of these variables were not available for the analyses summarised in Table 1: refusing consent at wave t might be followed by overall refusal at wave $t + 1$, and non-response might be greater in poorer neighbourhoods. The voting variable is an indicator of social engagement that might be related to the probability of responding. As the neighbourhood conditions score could not be obtained for cases that were not located, we use this variable just in the model that compares refusals with productives.

Table 2 presents the results using the same methods of estimation as for Table 1 with corresponding levels of precision. We see (from the notes) that each of the three variables is associated with at least one kind of non-response. The increase in accuracy of the AUC is more than would be expected by chance ($p < 0.001$ apart from wave non-response: $p > 0.06$) but is small except for refusal where the inclusion of the three new variables does make a difference: the estimate of the Gini coefficient increases

from 0.37 to 0.41 and the slope of the logit rank plot increases from 0.40 to 0.45 (although missing data for the neighbourhood conditions score does reduce the sample size).

4.2 Using weights to adjust for non-response

Although non-response at wave two of MCS is systematically related to a number of variables measured at or after wave one, we have seen that the models' ability to discriminate between and predict categories of non-response is not high. We now consider what effect the weights generated from the response propensity models have on a longitudinal estimate of interest. We focus on transitions between not working and working across the two waves. As Groves (2006) argues, the keys to unlocking missingness problems of bias are to find those variables that predict whether a piece of data is missing, and which of those variables that predict missingness are also related to the variable of interest. We find that all the variables that predict overall non-response are also related to whether or not the main respondent works at wave two, conditional on whether she was working at wave one so we might expect the application of non-response weights to reduce bias. The results are presented in Table 3 and show that, compared with just using the survey weights, the introduction of the non-response weights based on the model underpinning Table 1 leads to small adjustments in the estimated transition probabilities. The consent and vote variables have no additional effect, however, and this is consistent with the marginal increases in accuracy reported in Table 2.

Table 2
Accuracy estimates for enhanced response propensity models, MCS wave two

Accuracy measure	Overall non-response ⁽¹⁾	Non-response type		Non-response reason	
		Wave non-response ⁽²⁾	Attrition ⁽³⁾	Refusal ⁽⁴⁾	Other non-productive ⁽⁵⁾
AUC	0.70	0.72	0.71	0.70	0.77
Gini	0.41	0.44	0.41	0.41	0.54
Logit rank plot: slope	0.47	0.52	0.46	0.45	0.65
Sample size	18,148	16,177	16,745	15,656	16,443

⁽¹⁾ Includes consent (odds ratio (OR) = 2.1, s.e. = 0.20) and vote (OR = 1.4, s.e. = 0.08).
⁽²⁾ Includes vote only (OR = 1.4, s.e. = 0.11), consent not important ($t = 1.33$; $p > 0.18$).
⁽³⁾ Includes consent (OR = 2.7, s.e. = 0.26) and vote (OR = 1.4, s.e. = 0.09).
⁽⁴⁾ Includes consent (OR = 2.6, s.e. = 0.32), vote (OR = 1.3, s.e. = 0.10) and neighbourhood score (OR = 1.02, s.e. = 0.014).
⁽⁵⁾ Includes consent (OR = 1.6, s.e. = 0.20) and vote (OR = 1.5, s.e. = 0.11).

Table 3
Weighted employment transitions (standard errors), MCS wave two

Variable	Survey weights only	Overall weight ⁽¹⁾	Overall weight ⁽²⁾
No change	0.30 (0.0053)	0.30 (0.0056)	0.31 (0.0056)
Working → not working	0.34 (0.0059)	0.35 (0.0059)	0.35 (0.0060)
Not working → working	0.37 (0.0073)	0.35 (0.0073)	0.35 (0.0073)
Weight range ⁽³⁾	0.23 – 2.0	0.19 – 4.1	0.19 – 6.3
Sample size	14,891	14,796	14,733

⁽¹⁾ Based on the product of the survey weights and the non-response weights using the model underpinning Table 1.
⁽²⁾ Non-response weights based on a model that includes consent and vote.
⁽³⁾ All weights standardised to have mean of one.

5. Discussion

Survey methodologists working with longitudinal data have long been exercised by the problem of non-response. Nearly all longitudinal studies suffer from accumulating non-response over time and it is common even for well-conducted mature studies to obtain data for less than half the target sample. On the other hand, a lot can be learnt about the correlates of different types of non-response by drawing on auxiliary variables from earlier waves. The main purpose of this paper has been to introduce a different way of thinking about the utility of the approaches that rely on general linear models both to construct inverse probability weights and to inform imputations. Treating the linear predictors from the regression models as response propensity scores and then generating ROCs enables methods for summarising the information in these scores to be used to assess the accuracy of discrimination and prediction for different kinds of non-response.

The application of this approach to the Millennium Cohort Study has shown that, despite using a wide range of explanatory variables, discrimination is rather low. One implication of this finding is that some non-response is generated by circumstantial factors, none of them important on their own, which can reasonably be regarded as chance. There is some support for this hypothesis in that the accuracy of the models for overall non-response, wave non-response and other non-productive (the latter two being related) were little changed by the introduction of the voting and consent variables. On the other hand, these variables (and the neighbourhood conditions score) did improve the discrimination between productives, and attrition cases and refusals (which are also related). Nevertheless, discrimination for these two categories remained lower than for the other types of non-response. A second possible implication is that the models do not discriminate well because data are not missing at random (NMAR) in Little and Rubin's (2002) sense. In other words, it might be changes in circumstances after the previous wave that influences non-response at the current wave.

The implications of our findings for prediction are that it might be difficult to predict which cases will become non-respondents with a high degree of accuracy. If interventions to prevent non-response in longitudinal studies are to be effective then they need to be targeted at those cases least likely to respond because these cases are probably the most different from the respondents and therefore the major source of bias. This is where the ROC approach can be especially useful because, as Swets, Dawes and Monahan (2000) show, it is possible to determine the optimum threshold for the response propensity score based on the costs and benefits of intervening according to the true and

false positive rates implied by the threshold. A more detailed assessment of these issues is beyond the scope of this paper but would include considering interventions to prevent different kinds of non-response, and the benefits of potential reductions in bias and variability arising from a sample that is both larger and closer in its characteristics to the target sample.

Acknowledgements

This research was funded by the U.K. Economic and Social Research Council under its Survey Design and Measurement Initiative (ref. RES-175-25-0010).

References

- Copas, J. (1999). The effectiveness of risk scores: The logit rank plot. *Applied Statistics*, 48, 165-183.
- Groves, R.M. (2006). Nonresponse rates and non-response bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.
- Harrell, F.E. Jr., Lee, K.L. and Mark, D.B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361-387.
- Hawkes, D., and Plewis, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society A*, 169, 479-491.
- Krzanowski, W.J., and Hand, D.J. (2009). *ROC Curves for Continuous Data*. Boca Raton, FL: Chapman and Hall/CRC.
- Lepkowski, J.M., and Couper, M.P. (2002). Nonresponse in the second wave of longitudinal household surveys. In *Survey Nonresponse*, (Eds., R.M. Groves *et al.*). New York: John Wiley & Sons, Inc.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd Ed.). New York: John Wiley & Sons, Inc.
- Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: OUP.
- Plewis, I. (2007a). Non-response in a birth cohort study: The case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, 10, 325-334.
- Plewis, I. (Ed.) (2007b). *The Millennium Cohort Study: Technical Report on Sampling* (4th Ed.). London: Institute of Education, University of London.
- Plewis, I., Ketende, S.C., Joshi, H. and Hughes, G. (2008). The contribution of residential mobility to sample loss in a birth cohort study: Evidence from the first two waves of the Millennium Cohort Study. *Journal of Official Statistics*, 24, 365-385.
- Swets, J.A., Dawes, R.M. and Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Sciences in the Public Interest*, 1, 1-26.

Confidence interval estimation of small area parameters shrinking both means and variances

Sarat C. Dass, Tapabrata Maiti, Hao Ren and Samiran Sinha¹

Abstract

We propose a new approach to small area estimation based on joint modelling of means and variances. The proposed model and methodology not only improve small area estimators but also yield “smoothed” estimators of the true sampling variances. Maximum likelihood estimation of model parameters is carried out using EM algorithm due to the non-standard form of the likelihood function. Confidence intervals of small area parameters are derived using a more general decision theory approach, unlike the traditional way based on minimizing the squared error loss. Numerical properties of the proposed method are investigated via simulation studies and compared with other competitive methods in the literature. Theoretical justification for the effective performance of the resulting estimators and confidence intervals is also provided.

Key Words: EM algorithm; Empirical Bayes; Hierarchical models; Rejection sampling; Sampling variance; Small area estimation.

1. Introduction

Small area estimation and related statistical techniques have become a topic of growing importance in recent years. The need for reliable small area estimates is felt by many agencies, both public and private, for making useful policy decisions. An example where small area techniques are used in practice is in the monitoring of socio-economic and health conditions of different age-sex-race groups where the patterns are observed over small geographical areas.

It is now widely recognized that direct survey estimates for small areas are usually unreliable due to their typically large standard errors and coefficients of variation. Hence, it becomes necessary to obtain improved estimates with higher precision. Model-based approaches, either explicit or implicit, are elicited to connect the small areas and improved precision is achieved by “borrowing strength” from similar areas. The estimation technique is also known as shrinkage estimation since the direct survey estimates are shrunk towards the overall mean. The survey based direct estimates and sample variances are the main ingredients for building aggregate level small area models. The typical modeling strategy assumes that the sampling variances are known while a suitable linear regression model is assumed for the means. For details of these developments, we refer to reader to Ghosh and Rao (1994), Pfeffermann (2002) and Rao (2003). The typical area level models are subject to two main criticisms. First, in practice, the sampling variances are estimated quantities, and hence, are subject to substantial errors. This is because they are often based on equivalent sample sizes from which the direct estimates are calculated. Second, the assumption of known and fixed sampling variances of typical small area models does not take into

account the uncertainty in the variance estimation into the overall inference strategy.

Previous attempts have been made to model only the sampling variances; see, for example, Maples, Bell and Huang (2009), Gershunskaya and Lahiri (2005), Huff, Eltinge and Gershunskaya (2002), Cho, Eltinge, Gershunskaya and Huff (2002), Valliant (1987) and Otto and Bell (1995). The articles Wang and Fuller (2003) and Rivest and Vandal (2003) extended the asymptotic mean square error (MSE) estimation of small area estimators when the sampling variances are estimated as opposed to the standard assumption of known variances. Additionally, You and Chapman (2006) considered the modelling of the sampling variances with inference using full Bayesian estimation techniques.

The necessity of variance modelling has been felt by many practitioners. The latest developments in this area are nicely summarized in a recent article by William Bell of the United States Census Bureau 2008. He carefully examined the consequences of these issues in the context of MSE estimation of model based small area estimators. He also provided numerical evidence of MSE estimation for Fay-Herriot models (given in Equation 1) when sampling variances are assumed to be known. The developments in the small area literature so far can be “loosely” viewed as (i) smoothing the direct sampling error variances to obtain more stable variance estimates with low bias and (ii) (partial) accounting of the uncertainty in sampling variances by extending the Fay-Herriot model.

As evident, lesser or no attention has been given to account for the sampling variances effectively while modeling the mean compared to the volume of research that has been done for modeling and inferring the means. There is a lack of systematic development in the small area literature that

1. Sarat C. Dass and Tapabrata Maiti, Department of Statistics & Probability, Michigan State University. E-mail: maiti@stt.msu.edu; Hao Ren, CTB/McGraw-Hill, 20 Ryan Ranch Rd, Monterey, CA 93940; Samiran Sinha, Department of Statistics, Texas A & M University.

includes “shrinking” both means and variances. In other words, we like to exploit the technique of “borrowing strength” from other small areas to “improve” variance estimates as we do to “improve” the small area mean estimates. We propose a hierarchical model which uses both the direct survey and sampling variance estimates to infer all model parameters that determine the stochastic system. Our methodological goal is to develop the dual “shrinkage” estimation for both the small area means and variances, exploiting the structure of the mean-variance joint modelling so that the final estimators are more precise. Numerical evidence shows the effectiveness of dual shrinkage on small area estimates of the mean in terms of the MSE criteria.

Another major contribution of this article is to obtain confidence intervals of small area means. The small area literature is dominated by point estimates and their associated standard errors; it is well known that the standard practice of [point estimate $\pm q \times$ standard error], where q is the Z (standard normal) or t cut-off point, does not produce accurate coverage probabilities of the intervals; see Hall and Maiti (2006) and Chatterjee, Lahiri and Li (2008) for more details. Previous work is based on the bootstrap procedure and has limited use due to the repeated estimation of model parameters. We produce confidence intervals for the means from a decision theory perspective. The construction of confidence intervals is easy to implement in practice.

The rest of the article is organized as follows. The proposed hierarchical model for the sample means and variances is developed in Section 2. The estimation of model parameters via the EM algorithm is developed in Section 3. Theoretical justification for the proposed confidence interval and coverage properties are presented in Section 4. Sections 5 and 6 present a simulation study and a real data example, respectively. Some discussion and concluding remarks are presented in Section 7. An alternative model formulation for small area as well as mathematical details are provided in the Appendix.

2. Proposed model

Suppose n small areas are in consideration. For the i^{th} small area, let (X_i, S_i^2) be the pair of direct survey estimate and sampling variance, for $i = 1, 2, \dots, n$. Let $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})^T$ be the vector of p covariates available at the estimation stage for the i^{th} small area. We propose the following hierarchical model:

$$\left. \begin{aligned} X_i | \theta_i, \sigma_i^2 &\sim \text{Normal}(\theta_i, \sigma_i^2) \\ \theta_i &\sim \text{Normal}(\mathbf{Z}_i^T \boldsymbol{\beta}, \tau^2) \end{aligned} \right\} \quad (1)$$

$$\left. \begin{aligned} \frac{(n_i - 1)S_i^2}{\sigma_i^2} \Big| \sigma_i^2 &\sim \chi_{n_i-1}^2 \\ \sigma_i^{-2} &\sim \text{Gamma}(a, b), \end{aligned} \right\} \quad (2)$$

independently for $i = 1, 2, \dots, n$. In the model elicitation, n_i is the sample size for a simple random sample (SRS) from the i^{th} area, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is the $p \times 1$ vector of regression coefficients, and $\mathbf{B} \equiv (a, b, \boldsymbol{\beta}, \tau^2)^T$ is the collection of all unknown parameters in the model. Also, $\text{Gamma}(a, b)$ is the Gamma density function with positive shape and scale parameters a and b , respectively, defined as $f(x) = \{b^a \Gamma(a)\}^{-1} e^{-x/b} x^{a-1}$ for $x > 0$, and 0 otherwise. The unknown σ_i^2 is the true variance of X_i and is usually estimated by the sample variance S_i^2 . Although S_i^2 's are assumed to follow a chi-square distribution with $(n_i - 1)$ degrees of freedom (as a result of normality and SRS), we note that for complex survey designs, the degree of freedom needs to be determined carefully [e.g., Maples *et al.* 2009]. More importantly, the role of the sample sizes in shrinkage estimation of σ_i^2 is as follows: For low values of n_i , the estimate of σ_i^{-2} is shrunk more towards the overall mean (ab) compared to higher n_i values. Thus, for variances, sample sizes play the same role as precision in shrinkage estimation of the small area mean estimates. We note that You and Chapman (2006) also considered the second level of the sampling variance modelling. However, the hyperparameters related to prior of σ_i^2 are not data driven, they are rather chosen in such a way that the prior will be vague. Thus, their model can be viewed as the Bayesian version of the models considered in Rivest and Vandal (2003) and Wang and Fuller (2003). The second level modelling of σ_i^{-2} in (2) can be further extended to $\sigma_i^{-2} \sim \text{Gamma}(b, \exp(\mathbf{Z}_i^T \boldsymbol{\beta}_2)/b)$ so that $E(\sigma_i^{-2}) = \exp(\mathbf{Z}_i^T \boldsymbol{\beta}_2)$ for another set of p regression coefficients $\boldsymbol{\beta}_2$ to accommodate covariate information in the variance modelling.

Although our model is motivated by Hwang, Qiu and Zhao (2009), we like to mention that Hwang *et al.* (2009) considered shrinking means and variances in the context of microarray data where they prescribed an important solution by plugging in a shrinkage estimator of variance into the mean estimator. The shrinkage estimator of the variance in Hwang *et al.* (2009) is a function of S_i^2 only, and not of both X_i and S_i^2 ; see Remarks 2 and 3 in Section 2. Thus, inference of the mean does not take into account the full uncertainty in the variance estimation. Further, their model does not include any covariate information. The simulation study described subsequently indicate that our method of estimation performed better than Hwang *et al.* (2009).

In the above model formulation, inference for the small area mean parameter θ_i can be made based on the conditional distribution of θ_i given all of the data $\{(X_i, S_i^2, \mathbf{Z}_i), i = 1, \dots, n\}$. Under our model set up, the conditional

distribution of θ_i is a non-standard distribution and does not have a closed form, thus requiring numerical methods, such as Monte Carlo and the EM algorithm, for inference, and the details are provided in the next section.

3. Inference methodology

3.1 Estimation of unknown parameters via EM algorithm

In practice, $\mathbf{B} \equiv (a, b, \beta, \tau^2)^T$ is unknown and has to be estimated from the data $\{(X_i, S_i^2, \mathbf{Z}_i), i = 1, 2, \dots, n\}$. Our proposal is to estimate \mathbf{B} by the marginal maximum likelihood method: Estimate \mathbf{B} by $\hat{\mathbf{B}}$ where $\hat{\mathbf{B}}$ maximizes the marginal likelihood $L_M(\mathbf{B}) = \prod_{i=1}^n L_{M,i}(\mathbf{B})$, where

$$L_{M,i} \propto \frac{\Gamma(n_i/2+a)}{\tau \Gamma(a) b^a} \int \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2}\right\} \psi_i^{-(n_i/2+a)} d\theta_i, \quad (3)$$

and

$$\psi_i \equiv \left\{0.5(X_i - \theta_i)^2 + 0.5(n_i - 1)S_i^2 + \frac{1}{b}\right\}. \quad (4)$$

The marginal likelihood L_M involves integrals that cannot be evaluated in closed-form, and hence, one has to resort to numerical methods for its maximization. One such algorithm is the EM (Expectation-Maximization) iterative procedure which is used when such integrals are present. The EM algorithm involves augmenting the observed likelihood $L_M(\mathbf{B})$ with missing data; in our case, the variables of the integration, θ_i , $i = 1, 2, \dots, n$, constitute this missing information. Given $\theta \equiv \{\theta_1, \theta_2, \dots, \theta_n\}$, the complete data log likelihood (ℓ_c) can be written as

$$\ell_c(\mathbf{B}, \theta) = \sum_{i=1}^n \left[\log\{\Gamma(n_i/2+a)\} - \log\{\Gamma(a)\} - a \log(b) - 0.5 \log(\tau^2) - \frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2} - (n_i/2+a) \log(\psi_i) \right],$$

where the expression of ψ_i is given in Equation (4). Starting from an initial value of \mathbf{B} , $\mathbf{B}^{(0)}$ say, the EM algorithm iteratively performs a maximization with respect to \mathbf{B} . At the t^{th} step the objective function maximized is

$$\begin{aligned} Q(\mathbf{B} | \mathbf{B}^{(t-1)}) &= E(\ell_c(\mathbf{B}, \theta)) \\ &= \sum_{i=1}^n \left[\log\{\Gamma(n_i/2+a)\} - \log\{\Gamma(a)\} - a \log(b) - 0.5 \log(\tau^2) - \frac{E(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2} - (n_i/2+a) E\{\log(\psi_i)\} \right]. \end{aligned}$$

The expectation in $Q(\mathbf{B} | \mathbf{B}^{(t-1)})$ is taken with respect to the conditional distribution of each θ_i given the data, $\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}^{(t-1)})$, which is

$$\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \propto \exp\{-0.5(\theta_i - \mathbf{Z}_i^T \beta)^2 / \tau^2\} \psi_i^{-(n_i/2+a)}. \quad (5)$$

One challenge here is that the expectations are not available in closed form. Thus, we resort to a Monte Carlo method for evaluating the expressions. Suppose that R iid samples of θ_i are available, say $\theta_{i,1}, \theta_{i,2}, \dots, \theta_{i,R}$. Then, each expectation of the form $E\{h(\theta_i)\}$ can be approximated by the Monte Carlo mean

$$E\{h(\theta_i)\} \approx \frac{1}{R} \sum_{r=1}^R h(\theta_{i,r}). \quad (6)$$

However, drawing random numbers from the conditional distribution $\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}^{(t-1)})$ is also not straightforward since this is not a standard density. Samples are drawn using the accept-reject procedure (Robert and Casella 2004): For a sample from the target density f , sample x from the proposal density g , and accept the sample as a sample from f with probability $f(x)/\{M^*g(x)\}$ where $M^* = \sup_x \{f(x)/g(x)\}$. One advantage of the accept-reject method is that the target density f only needs to be known upto a constant of proportionality which is the case for $\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}^{(t-1)})$ in (5); due to the non-standard form of the density, the normalizing constant cannot be found in a closed form. For the accept-reject algorithm, we used the normal density $g(\theta_i) \propto \exp\{-0.5(\theta_i - \mathbf{Z}_i^T \beta)^2 / \tau^2\}$ as the proposal density. The acceptance probability is calculated to be $\{[1/b + 0.5(n_i - 1)S_i^2] / [1/b + 0.5(n_i - 1)S_i^2 + 0.5(\theta_i - X_i)^2]\}^{n_i/2+a}$. One can choose a better proposal distribution to increase acceptance probability or different algorithm (such as the adaptive rejection sampling or envelope accept-reject algorithms) but our chosen proposal worked satisfactorily in the studies we conducted.

The maximizer of $Q(\mathbf{B} | \mathbf{B}^{(t-1)})$ at the t^{th} step can be described explicitly. The solutions for β and τ^2 are available in closed form as

$$\beta^{(t)} = \left(\sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T \right)^{-1} \left(\sum_{i=1}^n \mathbf{Z}_i E(\theta_i) \right)$$

and

$$(\tau^2)^{(t)} = \frac{1}{n} \sum_{i=1}^n E(\theta_i - \mathbf{Z}_i^T \beta)^2,$$

respectively. Also, $a^{(t)}$ and $b^{(t)}$ are obtained by solving $S_a = \partial Q(\mathbf{B} | \mathbf{B}^{(t-1)}) / \partial a = 0$ and $S_b = \partial Q(\mathbf{B} | \mathbf{B}^{(t-1)}) / \partial b = 0$ using the Newton-Raphson method where

$$S_a = \sum_{i=1}^n \frac{\partial}{\partial a} \log\{\Gamma(n_i/2 + a)\} - n \left\{ \frac{\partial}{\partial a} \log\{\Gamma(a)\} \right\} - n \log(b) - \sum_{i=1}^n E\{\log(\psi_i)\}$$

and

$$S_b = -\frac{na}{b} + \sum_{i=1}^n \frac{(n_i/2 + a)}{b^2} E(\psi_i^{-1}).$$

We set $\mathbf{B}^{(t)} = (a^{(t)}, b^{(t)}, \boldsymbol{\beta}^{(t)}, (\tau^{(t)})^2)$ and proceed to the $(t+1)$ -st step. This maximization procedure is repeated until the estimate $\mathbf{B}^{(t)}$ converges. The MLE of \mathbf{B} , $\hat{\mathbf{B}} = \mathbf{B}^{(\infty)}$, once convergence is established.

3.2 Point estimate and confidence interval for θ_i

Following the standard technique, the small area estimator of θ_i is taken to be

$$\hat{\theta}_i = E(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \Big|_{\mathbf{B}=\hat{\mathbf{B}}}, \quad (7)$$

the expectation of θ_i with respect to the conditional density $\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$ with the maximum likelihood estimate $\hat{\mathbf{B}}$ plugged in for \mathbf{B} . The estimate $\hat{\theta}_i$ is calculated numerically using the Monte Carlo procedure (6) described in the previous section. Subsequently, all quantities involving the unknown \mathbf{B} will be plugged in by $\hat{\mathbf{B}}$ although we still keep using the notation \mathbf{B} for simplicity.

Further, we develop a confidence interval for θ_i based on a decision theory approach. Following Joshi (1969), Casella and Hwang (1991), Hwang *et al.* (2009), consider the loss function associated with the confidence interval C given by $(k/\sigma)L(C) - I_C(\theta)$ where k is a tuning parameter independent of the model parameters, $L(C)$ is the length of C and $I_C(\theta)$ is the indicator function taking values 1 or 0 depending on whether $\theta \in C$ or not. Note that this loss function takes into account both the coverage probability as well as the length of the interval; the positive quantity (k/σ) serves as the relative weight of the length compared to the coverage probability of the confidence interval. If $k = 0$, the length of the interval is not under consideration, which leads to the optimal C to be $(-\infty, \infty)$ with coverage probability 1. On the other hand, if $k = \infty$, then the coverage probability is 0, leading to optimal C to be a point set. The Bayes confidence interval for θ_i is obtained by minimizing the risk function (the expected loss) $E\{[(k/\sigma)L(C) - I_C(\theta)] | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}\}$. The optimal choice of C is given by

$$C_i(\mathbf{B}) = \{\theta_i: kE(\sigma_i^{-1} | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) < \pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})\}. \quad (8)$$

Since $C_i(\mathbf{B})$ is obtained by minimizing the posterior risk, one may like to interpret this as a Bayesian credible set. However, following Casella and Berger (1990, page 470), we will continue naming $C_i(\mathbf{B})$ as a confidence interval. From an empirical Bayes perspective also, this terminology is more appropriate. How the tuning parameter k determines the confidence level of $C_i(\mathbf{B})$ will be shown explicitly in Section 3.3.

Assuming k is known for the moment, we follow the steps below to calculate $C_i(\mathbf{B})$. The conditional densities of σ_i^2 and θ_i are given by

$$\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \propto \frac{\exp\left[\frac{-0.5(X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{(\sigma_i^2 + \tau^2) - \left\{0.5(n_i - 1)S_i^2 + \frac{1}{b}\right\}\left(\frac{1}{\sigma_i^2}\right)}\right]}{(\sigma_i^2)^{(n_i-1)/2+a+1}(\sigma_i^2 + \tau^2)^{1/2}} \quad (9)$$

and (5), respectively, which as mentioned before, are not available in closed form. Thus, similar to the case of θ_i , $E(\sigma_i^{-1} | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$ is computed numerically using the Monte Carlo method by approximating the expected value with the mean $1/N \sum_{k=1}^N 1/\sigma_{i,k}$ where $\sigma_{i,r}^2$, $r = 1, 2, \dots, R$ are R samples from the conditional density $\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$. The accept reject procedure is used to draw random numbers from $\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$ with a proposal density given by the inverse Gamma

$$\frac{\exp\left[-\left\{0.5(n_i - 1)S_i^2 + \frac{1}{b}\right\}\left(\frac{1}{\sigma_i^2}\right)\right]}{(\sigma_i^2)^{(n_i-1)/2+a+1}},$$

and the acceptance probability

$$\frac{\exp\left\{\frac{-0.5(X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{(\sigma_i^2 + \tau^2)}\right\}}{(\sigma_i^2 + \tau^2)^{1/2}} \times \exp(0.5) \times |X_i - \mathbf{Z}_i^T \boldsymbol{\beta}|.$$

The next step is to determine the boundary values of $C_i(\mathbf{B})$ by finding two θ_i values that satisfy the equation $kE(\sigma_i^{-1} | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) - \pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) = 0$. This requires the normalizing constant in (5)

$$D_i = \int_{-\infty}^{\infty} \exp\{-0.5(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 / \tau^2\} \psi_i^{-(n_i/2+a)} d\theta_i$$

to be evaluated numerically. This is obtained using the Gauss-Hermite integration with 20 nodes.

3.3 Choice of k

The choice of the tuning parameter k in (8) is taken to be

$$k = k(\mathbf{B}) = u_{i,0} \phi \left(t_{\alpha/2} \sqrt{\frac{n_i + 2a + 2}{n_i - 1}} \right) \quad (10)$$

where ϕ is the standard normal distribution, $t_{\alpha/2}$ is $(1 - \alpha/2)^{\text{th}}$ percentile of t distribution with $(n_i - 1)$ degrees of freedom, and $u_{i,0} = \sqrt{1 + \sigma_i^2/\tau^2}$. Since $u_{i,0}$ involves σ_i^2 which is unknown, an estimated version $\hat{u}_{i,0}$ is obtained by plugging in the maximum a posteriori estimate

$$\hat{\sigma}_i^2 = \hat{\sigma}_i^2(\hat{\mathbf{B}}) = \arg \max_{\sigma_i^2} \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \Big|_{\mathbf{B}=\hat{\mathbf{B}}} \quad (11)$$

in place of σ_i^2 . Also, \mathbf{B} is replaced by $\hat{\mathbf{B}}$ in (11). We demonstrate that the coverage probability of $C_i(\hat{\mathbf{B}})$ with this choice of k is close to $1 - \alpha$. Theoretical justifications are provided in Section 4.

3.4 Other related methods for comparison

Our method will be denoted as Method I. Three other methods to be compared are briefly described below.

Method II: Wang and Fuller (2003) considered the Fay-Herriot small area estimation model given by (1). Their primary contribution is the construction of the mean squared error estimation formulae for small area estimators with estimated sampling variances. In the process, they had constructed two formulae denoted by $\widehat{\text{MSE}}_1$ and $\widehat{\text{MSE}}_2$. We use $\widehat{\text{MSE}}_1$ for our comparisons, which was derived following the bias correction approach of Prasad and Rao (1990). The basic difference with our approach is that they did not smooth the sampling variances, only taking the uncertainty into account while making inference on the small area parameters. The method of parameter estimation, which is moment based for all the model parameters, is also different from ours.

Method III: Hwang *et al.* (2009) considered the log-normal and inverse Gamma models for σ_i^{-2} in (2) for microarray data analysis. Their simulation study showed improved performance of confidence intervals for small area estimators under the log-normal model compared to the inverse gamma. We thus modified their log-normal model to add covariates and for unequal sample sizes n_i as follows:

$$\left. \begin{aligned} X_i | \theta_i, \sigma_i^2 &\sim \text{Normal}(\theta_i, \sigma_i^2) \\ \theta_i &\sim \text{Normal}(\mathbf{Z}_i^T \boldsymbol{\beta}, \tau^2) \end{aligned} \right\} \quad (12)$$

$$\left. \begin{aligned} \log S_i^2 &= \log(\sigma_i^2) + \delta_i; \delta_i \sim N(m_i, \sigma_{ch,i}^2) \\ \log(\sigma_i^{-2}) &\sim N(\mu_v, \tau_v^2) \end{aligned} \right\} \quad (13)$$

independently for $i = 1, 2, \dots, n$. Note that the model for the means in (12) is identical to (1). The quantities τ^2 , m_i and $\sigma_{ch,i}^2$ are assumed to be known and are given by $m_i = E[\log(\chi_{n_i-1}^2/(n_i-1))]$ and $\sigma_{ch,i}^2 = \text{Var}[\log(\chi_{n_i-1}^2/(n_i-1))]$.

Thus, the sample size n_i 's determine the shape of the χ^2 distribution via its degrees of freedom parameter. More importantly, as mentioned earlier, the different sample sizes account for different degrees of shrinkage for the corresponding true variance parameter. Similar to their estimation approach, the unknown model parameters μ_v and τ_v^2 are estimated using a moment based approach in an empirical Bayes framework giving $\hat{\mu}_v$ and $\hat{\tau}_v^2$, respectively. Note that in Hwang *et al.* (2009), these estimates are obtained based on the hierarchical model for σ_i^2 of (13) *only* without regard to the modelling (1) of the mean. We refer to the Section 5 of their paper for details of the estimation of the hyper-parameters. We follow the same procedure using only (13) to estimate μ_v and τ_v^2 in the case of unequal sample sizes.

The Bayes estimate of σ_i^2 is derived to be

$$\begin{aligned} \hat{\sigma}_{i,B}^2 &= \exp[E\{\ln(\sigma_i^2) | \ln(S_i^2)\}] \\ &= \left\{ \frac{S_i^2}{\exp(m_i)} \right\}^{M_{v,i}} \exp\{\mu_v(1 - M_{v,i})\} \end{aligned}$$

where $M_{v,i} = \tau_v^2/(\tau_v^2 + \sigma_{ch,i}^2)$ and with estimates plugged in for the unknown quantities. The conditional distribution of θ_i given (X_i, S_i^2) , is

$$\pi(\theta_i | X_i, S_i^2) = \int_0^\infty \pi(\theta_i | X_i, S_i^2, \sigma_i^2) \pi(\sigma_i^2 | X_i, S_i^2) d\sigma_i^2,$$

is approximated as $\pi(\theta_i | X_i, S_i^2) \approx \int_0^\infty \pi(\theta_i | X_i, S_i^2, \hat{\sigma}_{i,B}^2) \pi(\sigma_i^2 | X_i, S_i^2) d\sigma_i^2 = \pi(\theta_i | X_i, S_i^2, \hat{\sigma}_{i,B}^2)$. This suggests the approximate Bayes estimator of the small area parameters given by

$$\hat{\theta}_i = E(\theta_i | X_i, \hat{\sigma}_{i,B}^2) = \hat{M}_i X_i + (1 - \hat{M}_i) \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}, \quad (14)$$

where $\hat{M}_i = \hat{\tau}_v^2/(\hat{\tau}_v^2 + \hat{\sigma}_{i,B}^2)$. The confidence interval for θ_i is obtained as

$$C_i^H = \left\{ \theta_i : \frac{|\theta_i - \hat{\theta}_i|}{\hat{M}_i \hat{\sigma}_{i,B}^2} < -2\ln\{k\sqrt{2\pi}\} - \ln(\hat{M}_i) \right\}. \quad (15)$$

In Section 3 of Hwang *et al.* (2009) pages 269-271, the interval C_i^H is matched with the $100(1 - \alpha)\%$ t -interval $[\theta_i - X_i | < t S_i]$ to obtain the expression of k as $k \equiv k_i = \exp\{-t^2/2\} \exp\{m_i/2\} / (\sqrt{2\pi})$.

Method IV: This method comprises of a special case of the Fay-Herriot model in (1) but with the estimation of model parameters adopted from Qiu and Hwang (2007). Qiu and Hwang (2007) considered the model

$$\left. \begin{aligned} X_i | \theta_i, \sigma^2 &\sim \text{Normal}(\theta_i, \sigma^2) \\ \theta_i &\sim \text{Normal}(0, \tau^2) \end{aligned} \right\} \quad (16)$$

independently for $i = 1, 2, \dots, n$, for analyzing microarray experimental data. When model parameters are known, they proposed the point estimator $\hat{\theta}_i = \hat{M}X_i$, $\hat{M} = (1 - ((n-2)\sigma^2/|X|^2))_+$ where a_+ denotes $\max(0, a)$ for any number a and $|X| = (\sum_{i=1}^n X_i^2)^{1/2}$. The confidence interval for θ_i is $\hat{\theta}_i \pm v_i(\hat{M})$, where $v_i^2(\hat{M}) = \sigma^2 \hat{M}(q_1 - \ln(\hat{M}))$ with q_1 denoting the standard normal cut-off point corresponding to desired level of confidence coefficient and $v_i(0) \equiv 0$. Here For the purpose of comparisons with our method, the first level of the hierarchical model in (16) is modified as follows:

$$X_i = \mathbf{Z}_i^T \boldsymbol{\beta} + v_i + e_i$$

where $v_i \sim \text{Normal}(0, \tau^2)$ and $e_i \sim \text{Normal}(0, S_i^2)$ independently for $i = 1, 2, \dots, n$, and S_i^2 is treated as known. Following Qiu and Hwang (2007), τ^2 is estimated by

$$\hat{\tau}^2 = \frac{1}{n-p} \left[\sum_i \hat{u}_i^2 - \sum_i S_i^2 \left\{ 1 - \mathbf{Z}_i^T \left(\sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T \right)^{-1} \mathbf{Z}_i^T \right\} \right]$$

and $\hat{\tau}^2 = \max(\hat{\tau}^2, 1/n)$ where $\hat{u}_i = X_i - \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T)^{-1} (\sum_{i=1}^n \mathbf{Z}_i X_i)$. Next, define $\hat{M}_{0i} = \hat{\tau}^2 / (\hat{\tau}^2 + S_i^2)$ and $\hat{M}_i = \max(\hat{M}_{0i}, M_1)$ where in the latter expression, \hat{M}_{0i} is truncated by $M_{1i} = 1 - Q_{\alpha} / (n_i - 2)$, and Q_{α} is the α^{th} quantile of a chi-squared distribution with n_i degrees of freedom. This \hat{M}_i is used in the formula of the confidence interval for θ_i given earlier. When applying this method in our simulation study and real data analysis, we modified the model to accommodate such unequal sample sizes and covariate information mentioned earlier.

Remark 1. Hwang *et al.* (2009) choose k by equating (15) to the t interval based on only X_i for the small area parameters θ_i . Note that X_i is the direct survey estimator. Consequently, this choice of k does not have any direct control over the coverage probability of the interval constructed under *shrinkage estimation*. On the other hand, our proposed choice of k has been derived to maintain nominal coverage under, specifically, shrinkage estimation.

Remark 2. Note that without any hierarchical modelling assumption, S_i and X_i are independent as S_i^2 and X_i are, respectively, ancillary and the complete sufficient statistics for θ_i . However, under models (1) and (2) the conditional distribution of σ_i^2 and θ_i involve both X_i and S_i^2 which is seen from (5) and (9).

Remark 3. In Hwang *et al.* (2009), the shrinkage estimator for σ_i^2 is based only on the information on S_i^2 , and not of both X_i and S_i^2 . The Bayes estimator of σ_i^2 is plugged into the expression for the Bayes estimator of small area parameters. Thus, Hwang *et al.*'s small area estimator is written as $E(\theta_i | X_i, \hat{\sigma}_{i,B}^2)$ in (14) where $\hat{\sigma}_{i,B}^2$ is the Bayes

estimator of σ_i^2 . Due to equation (9), the shrinkage estimator of σ_i^2 depends on $(X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2$ in addition to S_i^2 in contrast to Hwang *et al.* (2009). We believe this could be the reason for improved performance of our method compared to Hwang *et al.* (2009).

Remark 4. As mentioned previously, the degree of freedom associated with the χ^2 distribution for the sampling variance need not be simply $n_i - 1$, n_i being the sample size for i^{th} area. There is no sound theoretical result for determining the degree of freedom when the survey design is complex. The article Wang and Fuller (2003) approximated the χ^2 with a normal based on the Wilson-Hilferty approximation. If one knows the exact sampling design then the simulation based guideline of Maples *et al.* (2009) could be useful. For county level estimation using the American Community Survey, Maples *et al.* (2009) suggested the estimated degrees of freedom of $0.36 \times \sqrt{n_i}$.

4. Theoretical justification

Theoretical justification for the choice of k according to equation (10) is presented in this section. As in Hwang *et al.* (2009), the conditional distribution of θ_i given X_i and S_i^2 can be approximated as $\pi(\theta_i | X_i, S_i^2, \mathbf{B}) \approx \pi(\theta_i | X_i, S_i^2, \mathbf{B}, \hat{\sigma}_i^2)$, where $\hat{\sigma}_i^2$ as defined in (11). In a similar way, approximate $E(\sigma_i^{-1} | X_i, S_i^2, \mathbf{B})$ by $E(\sigma_i^{-1} | X_i, S_i^2, \mathbf{B}) \approx \hat{\sigma}_i^{-1}$. Based on these approximations, we have $C_i(\mathbf{B}) \approx \tilde{C}_i(\mathbf{B})$ where $\tilde{C}_i(\mathbf{B})$ is the confidence interval for θ_i given by $\tilde{C}_i(\mathbf{B}) = \{\theta_i: \pi(\theta_i | X_i, S_i^2, \mathbf{B}, \hat{\sigma}_i^2) \geq k \hat{\sigma}_i^{-1}\}$. From (1), it follows that the conditional density $\pi(\theta_i | X_i, S_i^2, \mathbf{B}, \hat{\sigma}_i^2)$ is a normal with mean μ_i and variance v_i , where μ_i and v_i are given by the expressions

$$\begin{aligned} \mu_i &= w_i X_i + (1 - w_i) \mathbf{Z}_i^T \boldsymbol{\beta}, \\ v_i &= \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right)^{-1} = \sigma_i^2 \left(1 + \frac{\sigma_i^2}{\tau^2} \right)^{-1}, \end{aligned} \quad (17)$$

and

$$w_i = \frac{1 / \sigma_i^2}{(1 / \sigma_i^2 + 1 / \tau^2)}.$$

Now, choosing

$$k = \hat{u}_0 \phi \left(t_{\alpha/2, \sqrt{\frac{n_i + 2a + 2}{n_i - 1}}} \right)$$

as discussed, the confidence interval $\tilde{C}_i(\mathbf{B})$ becomes

$$\tilde{C}_i(\mathbf{B}) = \left\{ \theta_i: \hat{u}_{0i} \frac{|\theta_i - \hat{\mu}_i|}{\hat{\sigma}_i} \leq t_{\alpha/2} \frac{n_i + 2a + 2}{n_i - 1} \right\}, \quad (18)$$

where $\hat{\mu}_i$ is the expression for μ_i in (17) with σ_i^2 replaced by $\hat{\sigma}_i^2$. Now consider the behavior of $\hat{\sigma}_i^2 \equiv \hat{\sigma}_i^2(\mathbf{B})$ as τ^2 ranges between 0 and ∞ . When $\tau^2 \rightarrow \infty$, $\hat{\sigma}_i^2$ converges to

$$\hat{\sigma}_i^2(\infty) \equiv \hat{\sigma}_i^2(a, b, \beta, \infty) = \frac{\frac{(n_i-1)S_i^2 + 1}{b}}{\frac{n_i-1}{2} + a+1} = \frac{(n_i-1)S_i^2 + \frac{2}{b}}{n_i + 2a + 1}.$$

Similarly, when $\tau^2 \rightarrow 0$, $\hat{\sigma}_i^2$ converges to

$$\hat{\sigma}_i^2(0) \equiv \hat{\sigma}_i^2(a, b, \beta, 0) = \frac{(X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 + (n_i-1)S_i^2 + \frac{2}{b}}{n_i + 2a + 2}.$$

For all intermediate values of τ^2 , we have $\min\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\} \leq \hat{\sigma}_i^2 \leq \max\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\}$. Therefore, it is sufficient to consider the following two cases: (i) $\hat{\sigma}_i^2 \geq \hat{\sigma}_i^2(\infty)$, where it follows that $(n_i + 2a + 2)\hat{\sigma}_i^2 = (n_i + 2a + 1)\hat{\sigma}_i^2 + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2 + 2/b + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2$, and (ii) $\hat{\sigma}_i^2 \leq \hat{\sigma}_i^2(0)$, where it follows that $(n_i + 2a + 2)\hat{\sigma}_i^2 = (X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 + (n_i - 1)S_i^2 + 2/b \geq (n_i - 1)S_i^2$. So, in both cases (i) and (ii),

$$(n_i + 2a + 2)\hat{\sigma}_i^2 \geq (n_i - 1)S_i^2. \quad (19)$$

Since $\theta_i - \mu_i \sim N(0, \sigma_i^2 \tau^2 / (\sigma_i^2 + \tau^2))$ and $(n_i - 1)S_i^2 / \sigma_i^2 \sim \chi_{n_i-1}^2$, the confidence interval

$$D_i = \left\{ \theta_i : u_{0i} \frac{|\theta_i - \mu_i|}{S_i} \leq t_{\alpha/2} \right\} \quad (20)$$

has coverage probability $1 - \alpha$. Thus, if u_0 and μ_i are replaced by \hat{u}_0 and $\hat{\mu}_i$, it is expected that the resulting confidence interval \tilde{D}_i , say, will have coverage probability of approximately $1 - \alpha$. From (19), we have

$$P(\tilde{C}_i(\mathbf{B})) \geq P(\tilde{D}_i) \approx 1 - \alpha, \quad (21)$$

establishing an approximate lower bound of $1 - \alpha$ for the confidence level of $\tilde{C}_i(\mathbf{B})$.

In (21), \mathbf{B} was assumed to be fixed and known. When \mathbf{B} is unknown, we replace \mathbf{B} by its marginal maximum likelihood estimate $\hat{\mathbf{B}}$. Since (21) holds regardless of the true value of \mathbf{B} , substituting $\hat{\mathbf{B}}$ for \mathbf{B} in (21) will involve an order $O(1/\sqrt{N})$ of error where $N = \sum_{i=1}^n n_i$. Compared to each single n_i , this pooling of n_i 's is expected to reduce the error significantly so that $\tilde{C}_i(\hat{\mathbf{B}})$ is sufficiently close to $\tilde{C}_i(\mathbf{B})$ to satisfy the lower bound of $1 - \alpha$ in (21).

5. A simulation study

5.1 Simulation setup

We considered a simulation setting using a subset of parameter configurations from Wang and Fuller (2003).

Each sample in the simulation study was generated from the following steps: First, generate observations using the model

$$X_{ij} = \beta + u_i + e_{ij},$$

where $u_i \sim N(0, \tau^2)$ and $e_{ij} \sim N(0, n_i \sigma_i^2)$, independently for $j = 1, \dots, n_i$ and $i = 1, \dots, n$. Then, the random effects model for the small area mean, X_i , is

$$X_i = \beta + u_i + e_i, \quad \text{independently for } i = 1, \dots, n,$$

where $X_i \equiv \bar{X}_i \equiv n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$ and $e_i \equiv \bar{e}_i \equiv n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$. Therefore, $X_i \sim N(\theta_i, \sigma_i^2)$ where $\theta_i = \beta + u_i$, $\theta_i \sim N(\beta, \tau^2)$ and $e_i \sim N(0, \sigma_i^2)$. We estimated σ_i^2 with the unbiased estimator

$$S_i^2 = (n_i - 1)^{-1} n_i^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,$$

and it follows that $(n_i - 1)S_i^2 / \sigma_i^2 \sim \chi_{n_i-1}^2$, independently for $i = 1, 2, \dots, n$. Note that the simulation layout has ignored the second level modeling of sampling variances in (2). Thus, our result will indicate robustness with respect to the variance model misspecification.

The above steps produced the data (X_i, S_i^2) , $i = 1, \dots, n$. To simplify the simulation, we do not choose any covariate information \mathbf{Z}_i . Similar to Wang and Fuller (2003), we set all n_i 's equal to m to ease programming efforts. However, the true sampling variances are still chosen to be unequal: One-third of the σ_i^2 are set to 1, another one-third are set to 4, and the remaining one-third are set to 16. We take $\beta = 10$ and three different choices of $\tau^2 = 0.25, 1$ and 4. These parameter values are chosen from Qiu and Hwang (2007). For each of τ^2 , we generated 200 samples for the two combinations $(m, n) = (9, 36)$ and $(18, 180)$.

In the simulation study, we compare the proposed method with the methods of Wang and Fuller (2003), Hwang *et al.* (2009) and Qiu and Hwang (2007) which are referred to as Methods I, II, III, and IV, respectively, based on bias, mean squared error (MSE), coverage probability (CP) of the confidence intervals and the length of the confidence intervals (ALCI). Table 1 contains the parameter estimates for a , b , β and τ^2 . The numerical results indicate good performance of the maximum likelihood estimates for the model parameters; the estimated values of β and τ^2 are close to the true values indicating good robustness properties with respect to distributional misspecification in the second level of (2). Statistically significant estimates for both a and b indicate that "shrunk" sampling variances are incorporated in the proposed method. Tables 2, 3 and 4 provide numerical results averaged over areas within each group having the same true sampling variances. The results in the Tables are based on 200 replications.

Table 1

Simulation results for the model parameters, a (top left panel), b (top right panel), β (bottom left panel) and τ^2 (bottom right panel). Here SD represents the standard deviation over 200 replicates. We took $\beta = 10$ and $\tau^2 = 0.25, 1$ and 4

τ^2	$n = 36, m = 9$			$n = 180, m = 18$			τ^2	$n = 36, m = 9$			$n = 180, m = 18$		
	Mean	SD		Mean	SD			Mean	SD		Mean	SD	
			a							b			
0.25	1.0959	0.1540		1.0328	0.0442		0.25	0.3992	0.0983		0.4249	0.0323	
1	1.0937	0.1555		1.0325	0.0445		1	0.4030	0.1012		0.4253	0.0326	
4	1.0996	0.1577		1.0339	0.0450		4	0.3999	0.1017		0.4245	0.0328	
			β							τ^2			
0.25	10.0071	0.3618		9.9951	0.1853		0.25	0.2558	0.0605		0.2575	0.0097	
1	10.0142	0.3311		9.9970	0.1743		1	0.9418	0.3333		1.0426	0.1264	
4	10.0282	0.4639		10.0048	0.2254		4	3.5592	1.3316		4.0817	0.5551	

Table 2

Simulation results for prediction when $\tau^2 = 0.25$. Here MSE, ALCI, CP represent the mean squared error, average confidence interval width, and coverage probability, respectively

	σ_i^2	$n = 36, m = 9$					$n = 180, m = 18$				
		Method					Method				
		I	II	III	IV		I	II	III	IV	
Relative bias	1	0.0048	0.0198	0.0272	0.0018		-0.0051	-0.0086	-0.0112	-0.0111	
	4	-0.0033	-0.0061	-0.0145	-0.0158		-0.0130	-0.0109	-0.0065	-0.0116	
	16	0.0126	0.0370	0.0369	0.0096		-0.0046	-0.0045	-0.0080	-0.0061	
MSE	1	0.3066	0.3890	0.6861	0.3805		0.2258	0.2680	0.4470	0.2922	
	4	0.3281	0.5430	1.3778	0.7285		0.2595	0.3000	0.5805	0.3748	
	16	0.3715	0.5240	1.6749	1.9316		0.2815	0.2850	0.4856	0.6383	
ALCI	1	2.1393	2.5485	4.4906	3.0528		1.9220	1.6006	3.6466	2.4811	
	4	2.2632	3.9574	6.8887	5.6842		2.0557	2.1524	5.2472	4.2160	
	16	2.3221	4.5619	9.3335	11.1363		2.1046	2.3308	6.5273	7.8492	
CP	1	0.9468	0.9770	0.9771	0.9708		0.9564	0.9710	0.9851	0.9631	
	4	0.9468	0.9710	0.9829	0.9917		0.9555	0.9660	0.9967	0.9967	
	16	0.9365	0.9660	0.9933	0.9975		0.9529	0.9610	0.9998	0.9999	

Table 3

Simulation results for prediction when $\tau^2 = 1$. Here MSE, ALCI, CP represent the mean squared error, average confidence interval width and coverage probability, respectively

	σ_i^2	$n = 36, m = 9$					$n = 180, m = 18$				
		Method					Method				
		I	II	III	IV		I	II	III	IV	
Relative bias	1	-0.0152	0.0205	0.0255	0.0051		-0.0064	-0.0085	-0.0111	-0.0101	
	4	-0.0167	-0.0164	-0.0151	-0.0219		-0.0151	-0.0121	-0.0133	-0.0164	
	16	-0.0323	0.0508	0.0515	0.0216		-0.0028	-0.0017	-0.0073	-0.0039	
MSE	1	0.5645	0.6330	0.7238	0.6260		0.5288	0.5430	0.5673	0.6336	
	4	0.8566	1.1100	1.5396	1.0992		0.8159	0.8770	0.9415	0.8948	
	16	1.0482	1.3100	2.1059	2.3156		0.9786	1.0000	1.1024	1.1878	
ALCI	1	3.4550	3.1822	4.4938	3.2117		3.1088	2.5094	3.6763	2.8676	
	4	4.0321	5.8733	6.8984	5.7909		3.7844	4.2908	5.3323	4.5543	
	16	4.4082	7.4286	9.3555	11.1555		4.1187	5.1590	6.6785	7.8937	
CP	1	0.9704	0.9640	0.9762	0.9275		0.9660	0.9650	0.9786	0.8879	
	4	0.9633	0.9560	0.9812	0.9808		0.9627	0.9680	0.9918	0.9740	
	16	0.9533	0.9490	0.9912	0.9938		0.9613	0.9680	0.9974	0.9979	

Table 4
Simulation results for prediction when $\tau^2 = 4$. Here MSE, ALCI, CP represent the mean squared error, average confidence interval length and the coverage probability, respectively

$n = 36, m = 9$						$n = 180, m = 18$			
Method						Method			
	σ_i^2	I	II	III	IV	I	II	III	IV
Relative bias	1	-0.0024	0.0248	0.0229	0.0180	-0.0084	-0.0098	-0.0122	-0.0106
	4	-0.0343	-0.0310	-0.0210	-0.0340	-0.0110	-0.0092	-0.0174	-0.0132
	16	-0.0147	0.0702	0.0767	0.0467	0.0016	0.0024	-0.0059	0.0012
MSE	1	0.8822	0.8590	0.8579	1.0559	0.8359	0.8180	0.8541	0.8605
	4	2.0577	2.2900	2.1818	2.2422	2.0424	2.1000	2.0935	2.1130
	16	3.4516	3.7600	3.9267	3.8981	3.3153	3.3500	3.3939	3.3631
ALCI	1	4.6318	4.1936	4.5369	3.7677	4.0256	3.5346	3.9626	3.7499
	4	6.2015	10.9093	7.0376	6.4314	5.9000	9.0913	6.2217	6.1540
	16	7.7221	18.0039	9.6718	11.3341	7.4430	14.6665	8.3908	8.7537
CP	1	0.9791	0.9670	0.9733	0.9029	0.9674	0.9570	0.9600	0.9468
	4	0.9556	0.9670	0.9725	0.9496	0.9592	0.9610	0.9633	0.9573
	16	0.9510	0.9670	0.9796	0.9858	0.9573	0.9650	0.9718	0.9776

Bias Comparisons: In most cases, the bias of the four methods are comparable. There is no clear evidence of significant differences between them in terms of the bias. High sampling variance gives more weight to the population mean by construction that makes the estimator closer to the mean at the second level. On the other hand, Methods I - III use shrinkage estimators of the sampling variances which would be less than the maximum of all sampling variances. Thus, Methods I - III tend to have little more bias. However, due to shrinkage in sampling variances, one may expect a gain in the variance of the estimators which, in turn, makes the MSE smaller. Among Methods I - III, Method I performed better compared to Methods II and III, which were quite similar to each other. The maximum gain using Method I compared to Method II is 99%.

MSE Comparisons: In terms of the MSE, Method I performed consistently better than the other three in all cases except when the ratio of σ_i^2 to τ^2 is the lowest: ($\sigma_i^2 = 1$) / ($\tau^2 = 4$) = 0.25. In this case, the variance between small areas (model variance) is much higher than the variance within the areas (sampling variance). When using our method to estimate θ_i , the information “borrowed” from other areas may misdirect the estimation: The estimated mean of the Gamma distribution for σ_i^{-2} from the second level in (2) is $\hat{a}\hat{b}$ which equals 0.44 approximately for both the (m, n) combinations of (9, 36) and (18, 180) (the true value is $ab = 0.4$). Thus, $E(\sigma_i^{-2} | X_i, S_i^2, \hat{B})$ is significantly smaller than 1 due to shrinkage towards the mean for the group which has the true value of $\sigma_i^2 = 1$. Also, since σ_i^2 is smaller than τ^2 , the weight of X_i should be much more compared to β , the overall mean. However,

due to underestimation of σ_i^{-2} in this case, the resulting estimator puts less weight on X_i which leads to higher MSE. However, this underestimation will decrease for large sample sizes due to the consistency of Bayes estimators. This fact is actually observed when the sample size increases from $n = 36$ to $n = 180$ for the case $\sigma_i^2 = 1$ and $\tau^2 = 4$. Compared to Method II, Method I shows gains in most of the simulation cases; the maximum gain is 30% while the only loss is 9% for the combination $\sigma_i^2 = 1$ and $\tau^2 = 4$ for $n = 36$ and $m = 9$. Similarly, for Method III, the maximum gain of Method I is 77% and the only loss of 11% is for the same parameter and sample size specifications.

ACP Comparisons: We obtained confidence intervals with confidence level 95%. Methods I and III do not indicate any under-coverage. This is expected from their optimal confidence interval construction. Method I meets the nominal coverage rate more frequently than any other methods. Method II has some under coverage and can go as low as 82%.

ALCI Comparisons: Method I produced considerably shorter confidence intervals in general. Method IV produced comparable lengths as the other methods in all cases except when σ_i^2 was high, in which case, the lengths were considerably higher. The confidence interval proposed in Qiu and Hwang (2007) does not have good finite sample properties, particularly for small τ^2 . To avoid low coverage, they proposed to truncate $M_0 = \tau^2 / (\tau^2 + \sigma_i^2)$ with a positive number $M_1 = 1 - Q_\alpha / (v - 2)$ for known σ_i^2 where Q_α is the α^{th} -quantile of a chi-squared distribution with v degrees of freedom. When the ratio of sampling

variance to model variance, σ_i^2/τ^2 , is high, M_1 tends to be higher than M_0 . This results in a nominal coverage but with larger interval lengths. For example, in case of $(\sigma_i^2, \tau^2) = (16, 0.25)$, the ALCI is 11.13 for Method IV whereas ALCI is only 2.78 and 4.56 for Methods I and II.

5.2 Robustness study

In order to study the robustness of the proposed method with respect to departures from the normality assumption in the errors, we conducted the following simulation study. Data was generated as before but with e_{ij} 's drawn from a double-exponential (Laplace) and an uniform distribution. The estimators from Methods II and III had little effect. This is perhaps due to the fact that these methods used moment based estimation for model parameter estimation. Method IV resulted in larger relative bias, MSE and ALCI, and lower coverage probability. The MSE from Method I is always lower than that from Method II. For $\tau^2 = 0.25$ and 1, ALCI is smaller for Method I compared to Method II for $(n = 36, m = 9)$ but the results are opposite when $(n = 180, m = 18)$. In terms of CP, Method II has some under coverage (lowest is 80%). However, Method I did not have any under-coverage. In order to save space we only provide

the results for parameters a, b, β and τ^2 under the Laplace errors (see Table 5).

6. Real data analysis

We illustrate our methodology based on a widely studied example. The data set is from the U.S. Department of Agriculture and was first analyzed by Battese (1988). The data set is on corn and soybeans productions in 12 Iowa counties. The sample sizes for these areas are small, ranging from 1 to 5. We shall consider corn only to save space. For the proposed model, the sample sizes $n_i > 1$ necessarily. Therefore, modified data from You and Chapman (2006) with $n_i \geq 2$ are used. The mean reported crop hectares for corn (X_i) are the direct survey estimates and are given in Table 6. Table 6 also gives the sample variances which are calculated based on the original data assuming simple random sampling. The sample standard deviation varies widely, ranging from 5.704 to 53.999 (the coefficient of variation varies from 0.036 to 0.423). Two covariates are considered in Table 6: Z_{1i} , the mean of pixels of corn, and Z_{2i} , the mean of pixels of soybean, from the LANDSAT satellite data.

Table 5
Simulation results for the model parameters, a (top left panel), b (top right panel), β (bottom left panel) and τ^2 (bottom right panel) when the errors follow a laplace distribution. Here SD represents the standard deviation over 200 replicates. We took $\beta = 10$ and $\tau^2 = 0.25, 1$ and 4

τ^2	$n = 36, m = 9$		$n = 180, m = 18$		τ^2	$n = 36, m = 9$		$n = 180, m = 18$	
	Mean	SD	Mean	SD		Mean	SD	Mean	SD
a					b				
0.25	0.9624	0.1632	0.9471	0.0498	0.25	0.5793	0.1733	0.5279	0.0501
1	0.9628	0.1657	0.9476	0.0497	1	0.5816	0.1777	0.5275	0.0503
4	0.9689	0.1694	0.9487	0.0499	4	0.5758	0.1796	0.5263	0.0503
β					τ^2				
0.25	9.9736	0.3775	9.9800	0.1773	0.25	0.2696	0.0882	0.2565	0.0074
1	9.9753	0.3709	9.9836	0.1662	1	1.0508	0.2501	1.0403	0.0668
4	9.9736	0.4835	9.9855	0.2161	4	3.9624	1.1719	4.1256	0.4201

Table 6
Corn data from You and Chapman (2006)

County	n_i	X_i	Z_{1i}	Z_{2i}	$\sqrt{S_i^2}$
Franklin	3	158.623	318.21	188.06	5.704
Pocahontas	3	102.523	257.17	247.13	43.406
Winnebago	3	112.773	291.77	185.37	30.547
Wright	3	144.297	301.26	221.36	53.999
Webster	4	117.595	262.17	247.09	21.298
Hancock	5	109.382	314.28	198.66	15.661
Kossuth	5	110.252	298.65	204.61	12.112
Hardin	5	120.054	325.99	177.05	36.807

The estimates of \mathbf{B} are as follows: $a = 1.707$, $b = 0.00135$, $\tau^2 = 90.58$ and $\beta = (-186.0, 0.7505, 0.4100)$. The estimated prior mean of $1/\sigma_i^2$ which is the mean of the Gamma distribution with parameters a and b is $ab = 0.002295$ with a square root of 0.048 (note that $1/0.048 = 20.85$ consistent with the range of the sample standard deviations between 5.704 and 53.999). The small area estimates and their confidence intervals are summarized in Table 7 and Figure 1. Point estimates of all 4 methods are comparable: the summary measures comprising of the mean, median, and range of the small area parameter estimates for Methods I, II, III, and IV are (121.9, 124.1, 122.2, 122.6), (125.2, 120.4, 115.0, 114.5) and (23.1, 53.0, 58.4, 56.6), respectively. The distribution of $\hat{\theta}_i$ (plotted based on considering all the i 's) are summarized in Figure 2 which shows that there is a significant difference in their variability. Method I has the lowest variability and is superior in this sense. Further, smoothing sampling variances has strong implication in measuring uncertainty and hence in the interval estimation. The proposed method has the shortest confidence interval on an average compared to all other methods. Methods II and III provide intervals with negative lower limits. This seems unrealistic because the direct average of area under corn is positive and large for all the 12 counties (the crude confidence intervals $(x_i \pm t_{0.025} S_i)$ do not contain zero for any of the areas either). Note that Method II does not have any theoretical support on its confidence intervals. Methods II and III produce wider confidence intervals when the sampling variance is high. For example, the sample size for both Franklin county

and Pocahontas county is three, but sampling standard deviations are 5.704 and 43.406. Although the confidence interval under Method I is comparable, they are wide apart for Methods II and III. This is because although these methods consider the uncertainty in sampling variance estimates, the smoothing did not use the information from direct survey estimates, resulted the underlying sampling variance estimates remain highly variable (due to small sample size). In effect, the variance of the variance estimator (of the point estimates) is bigger compared to that in method I. This is further confirmed by the fact that the intuitive standard deviations of the “smoothed” small area estimates (one fourth of the interval) are smaller and less variable under method I compared to the others. Another noticeable aspect of our method is that the interval widths are similar for counties with same sample size. This could be an indication of obtaining equ-efficient estimators for equivalent sample sizes.

Model selection: For choosing the best fitting model, we used the Bayesian Information Criteria (BIC) which takes into account both the likelihood as well as the complexity of the fitted models. We calculated BICs for the models used in Methods I and III (Hwang *et al.* 2009). These two models have the same numbers of parameters with a difference in only the way the parameters are estimated. The model BIC for Method I is 210.025 and that for Method III is 227.372. This indicates superiority of our model. We could not compute the BIC for Wang and Fuller (2003) since they did not use any explicit likelihood.

Table 7
Results of the corn data analysis. Here CI and LCI represent the confidence interval and the length of the confidence interval, respectively

County	$\hat{\theta}_i$	CI	LCI	$\hat{\theta}_i$	CI	LCI
I: Proposed method				II: Wang and Fuller (2003)		
Franklin	131.8106	104.085, 159.372	55.287	155.4338	124.151, 193.094	68.943
Pocahontas	108.7305	80.900, 136.436	55.536	102.3682	-38.973, 244.019	282.993
Winnebago	109.0559	81.430, 136.646	55.216	115.9093	-53.768, 279.314	333.083
Wright	131.6113	103.736, 159.564	55.828	131.0674	8.330, 280.263	271.932
Webster	113.1484	92.805, 133.348	40.543	109.4795	32.514, 202.675	170.161
Hancock	129.4279	111.781, 147.193	35.412	124.1028	56.750, 162.013	105.262
Kossuth	121.0071	103.451, 138.626	35.175	116.7147	68.049, 152.454	84.405
Hardin	130.2520	112.373, 148.114	35.741	137.7983	51.734, 188.373	136.638
III: Hwang <i>et al.</i> (2009)				IV: Qiu and Hwang (2007)		
Franklin	158.4677	128.564, 188.370	59.805	157.7383	146.999, 168.477	21.478
Pocahontas	100.1276	-44.039, 244.295	288.334	101.1661	19.444, 182.887	163.442
Winnebago	114.1473	0.065, 228.228	228.163	113.7746	56.263, 171.286	115.022
Wright	140.3717	-24.119, 304.862	328.982	143.2244	41.559, 244.889	203.330
Webster	115.7865	50.297, 181.275	130.978	115.2224	75.124, 155.320	80.196
Hancock	111.3087	66.213, 156.403	90.189	113.1766	83.691, 142.661	58.970
Kossuth	110.9585	74.366, 147.550	73.184	112.3239	89.520, 135.127	45.607
Hardin	126.6093	40.040, 213.178	173.137	123.9049	54.607, 193.202	138.594

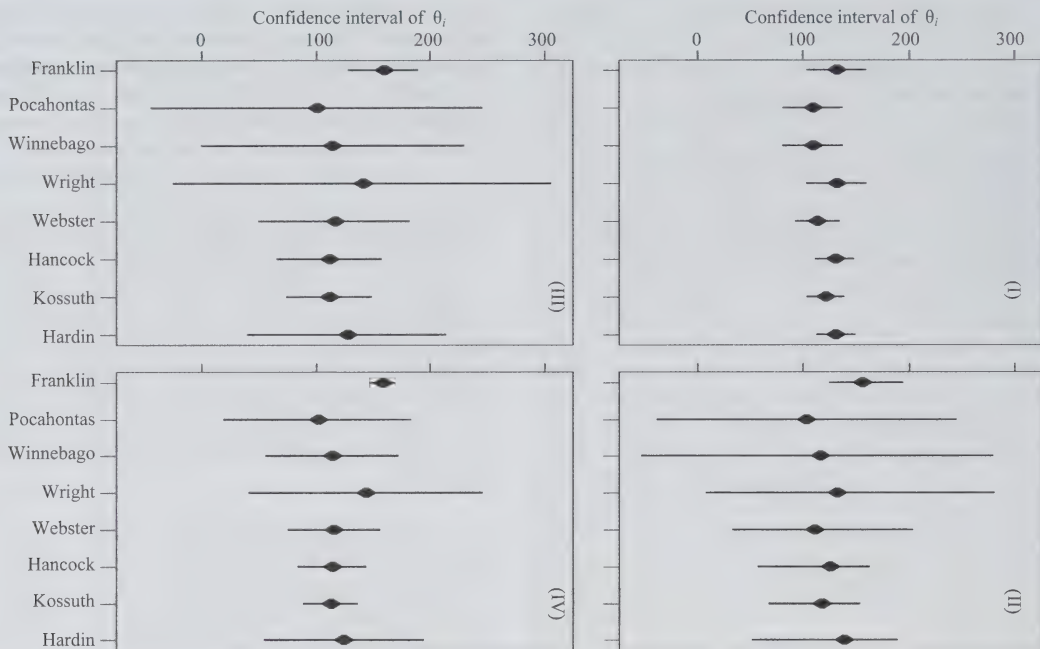


Figure 1 Corn hectares estimation. The horizontal line for each county displays the confidence interval of $\hat{\theta}_i$, with $\hat{\theta}_i$ marked by the circle, for (I) Proposed method, (II) Wang and Fuller (2003), (III) Hwang *et al.* (2009) and (IV) Qiu and Hwang (2007)

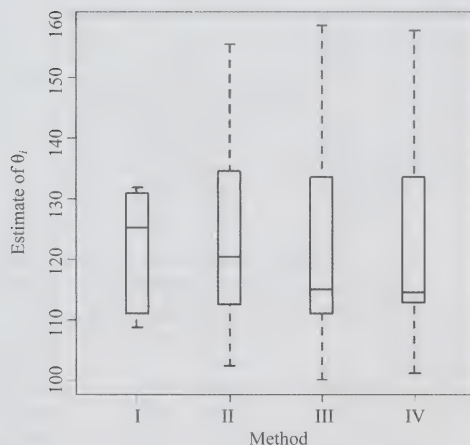


Figure 2 Boxplot of estimates of corn hectares for each county. (I) to (IV) are the 4 methods corresponding to Figure 1

7. Conclusion

In this paper, joint area level modeling of means and variances is developed for small area estimation. The resulting small area estimators are shown to be more efficient than the traditional estimators obtained using Fay-Herriot models which only shrink the means. Although our model is same as one considered in Hwang *et al.* (2009), our method of estimation is different in two ways: In the determination of the tuning parameter k and the use of $\pi(\sigma_i^2 | X_i, S_i^2, Z_i)$ (which depends additionally on X_i), instead of $\pi(\sigma_i^2 | S_i^2, Z_i)$, for constructing the conditional distribution of the small area parameters θ_i . We demonstrated robustness properties of the model when the assumption that σ_i^2 arise from a inverse Gamma distribution is violated. The borrowing of X_i information when estimating σ_i^2 as well as the robustness with respect to prior elicitation demonstrate the superiority of our proposed method. The parameter values chosen in the simulation study are different than in the real data analysis. The real data analysis given here is merely for illustration purposes. Our main aim was to

develop the methodology of mean-variance modeling and contrast with some closely related methods to show its effectiveness. For this reason, we chose parameter settings in the simulation to be the same as in the well-known small area estimation article Wang and Fuller (2003).

Obtaining improved sampling variance estimators is a byproduct of the proposed approach. We have provided an innovative estimation technique which is theoretically justified and user friendly. Computationally, the method is much simpler compared to other competitive methods such as Bayesian MCMC procedures or bootstrap resampling methods. We need sampling from posterior distribution only once during the model parameter estimation, and the sampled values can be used subsequently for all other purposes. The software is available from the authors upon request.

Acknowledgements

The authors like to thank two referees and the Associate Editor for their constructive comments that have led to a significantly improved version of this article. The research is partially supported by NSF grants SES 0961649, 0961618 and DMS 1106450.

Appendix

A. Derivation of the conditional distributions

From Equation (1) and (2), the conditional joint distribution of $\{X_i, S_i^2, \theta_i, \sigma_i^2\}$, $\pi(X_i, S_i^2, \theta_i, \sigma_i^2 | a, b, \beta, \tau^2)$, is

$$\begin{aligned} \pi(X_i, S_i^2, \theta_i, \sigma_i^2 | \mathbf{Z}_i, \mathbf{B}) &= \frac{1}{2\pi\sigma_i^2} \exp\left\{-\frac{(X_i - \theta_i)^2}{2\sigma_i^2}\right\} \frac{1}{\Gamma\left(\frac{n_i-1}{2}\right) 2^{\frac{n_i-1}{2}}} \\ &\times \left\{(n_i-1) \frac{S_i^2}{\sigma_i^2}\right\}^{\frac{n_i-1}{2}-1} \exp\left\{-\frac{(n_i-1)S_i^2}{2\sigma_i^2}\right\} \\ &\times \left(\frac{n_i-1}{\sigma_i^2}\right) \frac{1}{2\pi\tau^2} \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2}\right\} \\ &\times \frac{1}{\Gamma(a)b^a} \left(\frac{1}{\sigma_i^2}\right)^{a+1} \exp\left(-\frac{1}{b\sigma_i^2}\right) \\ &\propto \exp\left[-\frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2} - \left\{\frac{(X_i - \theta_i)^2}{2} + \frac{(n_i-1)S_i^2}{2} + \frac{1}{b}\right\} \frac{1}{\sigma_i^2}\right] \\ &\times \left(\frac{1}{\sigma_i^2}\right)^{\frac{n_i}{2}+a+1} \left(\frac{1}{\tau^2}\right)^{\frac{1}{2}} \frac{1}{\Gamma(a)b^a}. \end{aligned}$$

Therefore the conditional distribution of σ_i^2 and θ_i given the data and \mathbf{B} are

$$\begin{aligned} \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) &= \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 | \mathbf{Z}_i, \mathbf{B}) d\theta_i \propto \frac{1}{(\sigma_i^2)^{(n_i-1)/2+a+1} (\sigma_i^2 + \tau^2)^{1/2}} \\ &\exp\left[-\frac{(X_i - \mathbf{Z}_i^T \beta)^2}{2(\sigma_i^2 + \tau^2)} - \left\{\frac{1}{2}(n_i-1)S_i^2 + \frac{1}{b}\right\} \left(\frac{1}{\sigma_i^2}\right)\right], \\ \pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) &= \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 | \mathbf{Z}_i, \mathbf{B}) d\sigma_i^2 \\ &\propto \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} \end{aligned}$$

where ψ_i is defined in Equation (4).

B. Details of the EM algorithm

The maximization of $Q(\mathbf{B} | \mathbf{B}^{(r-1)})$ is done by setting the partial derivatives with respect to \mathbf{B} to be zero, that is,

$$\frac{\partial Q(\mathbf{B} | \mathbf{B}^{(r-1)})}{\partial \mathbf{B}} = 0. \quad (\text{B.1})$$

From the expression of $Q(\mathbf{B} | \mathbf{B}^{(r-1)})$ in the text, we give explicit expressions for the partial derivatives with respect to each component of \mathbf{B} . The partial derivative corresponding to β is

$$\begin{aligned} \frac{\partial Q(\mathbf{B} | \mathbf{B}^{(r-1)})}{\partial \beta} &= \sum_{i=1}^n \frac{\int \mathbf{Z}_i \left(\frac{\theta_i - \mathbf{Z}_i^T \beta}{\tau^2}\right) \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i}{\int \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i} \\ &= \sum_{i=1}^n E\left\{\mathbf{Z}_i \left(\frac{\theta_i - \mathbf{Z}_i^T \beta}{\tau^2}\right)\right\} \end{aligned}$$

where the expectation is with respect to the conditional distribution of θ_i , $\pi(\theta_i | X_i, S_i^2, \mathbf{B})$. The expression of the partial derivative corresponding to τ^2 is:

$$\begin{aligned} \frac{\partial Q(\mathbf{B} | \mathbf{B}^{(r-1)})}{\partial \tau^2} &= -\frac{n}{2\tau^2} + \sum_{i=1}^n \frac{\int \frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2(\tau^2)^2} \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i}{\int \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i} \\ &= -\frac{n}{2\tau^2} + \sum_{i=1}^n E\left\{\frac{(\theta_i - \mathbf{Z}_i^T \beta)^2}{2(\tau^2)^2}\right\}. \end{aligned}$$

Similarly for a and b , we get the solutions by setting $S_a = 0$ and $S_b = 0$ where S_a and S_b are, respectively, the partial derivatives of $Q(\mathbf{B}|\mathbf{B}^{(t-1)})$ with respect to a and b with expressions given in the main text. These equations are solved using the Newton-Raphson method which requires the matrix of second derivatives with respect to a and b . These are given by the following expressions:

$$\begin{aligned} S_{aa} &= \sum_{i=1}^n \left[\log'' \left\{ \Gamma \left(\frac{n_i}{2} + a \right) \right\} \right. \\ &\quad \left. - \log'' \{ \Gamma(a) \} + \text{Var} \{ \log(\psi_i) \} \right] \\ S_{ab} &= \sum_{i=1}^n \left[-\frac{1}{b} + \frac{1}{b^2} E \left(\frac{1}{\psi_i} \right) - \left(\frac{n_i}{2} + a \right) \frac{1}{b^2} \right. \\ &\quad \left. \text{Cov} \left\{ \frac{1}{\psi_i}, \log(\psi_i) \right\} \right], \\ \text{and} \\ S_{bb} &= \sum_{i=1}^n \left\{ \frac{a}{b^2} - (n_i + 2a) \frac{1}{b^3} E \left(\frac{1}{\psi_i} \right) + \left(\frac{n_i}{2} + a \right) \frac{1}{b^4} \right. \\ &\quad \left. E \left(\frac{1}{\psi_i^2} \right) + \left(\frac{n_i}{2} + a \right)^2 \frac{1}{b^4} \text{Var} \left(\frac{1}{\psi_i} \right) \right\} \end{aligned} \quad (\text{B.2})$$

with $S_{ba} = S_{ab}$. At the u^{th} step, the update of a and b are given by

$$\begin{bmatrix} a^{(u)} \\ b^{(u)} \end{bmatrix} = \begin{bmatrix} a^{(u-1)} \\ b^{(u-1)} \end{bmatrix} - \begin{bmatrix} S_{aa}^{(u-1)} & S_{ab}^{(u-1)} \\ S_{ba}^{(u-1)} & S_{bb}^{(u-1)} \end{bmatrix}^{-1} \begin{bmatrix} S_a^{(u-1)} \\ S_b^{(u-1)} \end{bmatrix}, \quad (\text{B.3})$$

where the superscript $(u-1)$ on S_{aa} , S_{ab} , S_{ba} , S_{bb} , S_a and S_b denote these quantities evaluated at the values of a and b at the $(u-1)^{\text{th}}$ iteration. Once the Newton Raphson procedure converges, the value of a and b at the t^{th} step of the EM algorithm is set as $a^{(t)} = a^{(\infty)}$ and $b^{(t)} = b^{(\infty)}$.

C. An alternative small area model formulation

It is possible to reduce the width of the confidence interval $\bar{C}(\mathbf{B})$ based on an alternative hierarchical model for small area estimation which has some mathematical elegance. The constant term $n_i + 2a + 2$ in (19) becomes $n_i + 2a$ in this alternative model formulation. The model is given by

$$X_i | \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \quad (\text{C.1})$$

$$\theta_i | \sigma_i^2 \sim N(\mathbf{Z}_i \boldsymbol{\beta}, \lambda \sigma_i^2), \quad (\text{C.2})$$

$$\frac{(n_i - 1) S_i^2}{\sigma_i^2} \Big| \sigma_i^2 \sim \chi_{n_i - 1}^2, \quad (\text{C.3})$$

$$\sigma_i^2 \sim \text{Inverse - Gamma}(a, b), \quad (\text{C.4})$$

independently for $i = 1, 2, \dots, n$. Note that in the above formulation, it is assumed that the conditional variance of θ_i is proportional to σ_i^2 whereas the marginal variance is constant (by integrating out σ_i^2 using (C.4). In (1) and (2), the variance of θ_i is a constant, τ^2 , independent of σ_i^2 , and there is no conditional structure for θ_i depending on σ_i^2 . The set of all unknown parameters in the current hierarchical model is $\mathbf{B} = (a, b, \boldsymbol{\beta}, \lambda)$. The inference procedure for this model is given subsequently. The model essentially assumes that the true small area effects are not identically distributed even after eliminating the known variations.

C.1 Inference methodology

By re-parameterizing the variance as in (C.2), some analytical simplifications are obtained in the derivation of the posteriors of θ_i and σ_i given X_i , S_i^2 and \mathbf{B} . We have

$$\begin{aligned} \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{B}) \\ = IG \left(\frac{n_i}{2} + a, \left[\frac{(n_i - 1) S_i^2}{2} + \frac{(X_i - \mathbf{Z}_i \boldsymbol{\beta})^2}{2(1 + \lambda)} + \frac{1}{b} \right] \right) \end{aligned}$$

where $IG(a, b)$ stands for the inverse Gamma distribution with shape and scale parameters a and b , respectively. Given \mathbf{B} and σ_i^2 , the conditional distribution of θ_i is

$$\pi(\theta_i | X_i, \sigma_i^2, \mathbf{B}) = \text{Normal} \left(\mathbf{Z}_i^T \boldsymbol{\beta}, \frac{\lambda \sigma_i^2}{1 + \lambda} \right).$$

Integrating out σ_i^2 , one obtains the conditional distribution of θ_i given X_i , S_i^2 and \mathbf{B} ,

$$\begin{aligned} \pi(\theta_i | X_i, S_i^2, \mathbf{B}) \\ = \int_0^\infty \pi(\theta_i | X_i, \sigma_i^2, \mathbf{B}) \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{B}) d\sigma_i^2 \\ \propto \left\{ \frac{(1 + \lambda)}{2\lambda} (\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 + \frac{\delta^2}{2} \right\}^{-(n_i + 2a + 1)/2}, \quad (\text{C.5}) \end{aligned}$$

where $\delta^2 = (n_i - 1) S_i^2 + (X_i - \mathbf{Z}_i \boldsymbol{\beta})^2 / (1 + \lambda) + 2/b$. We can rewrite (C.5) as

$$\begin{aligned} \pi(\theta_i | X_i, S_i^2, \mathbf{B}) &= \frac{\Gamma((n_i + 1)/2 + a) \sqrt{1 + \lambda}}{\delta^* \Gamma(n_i/2 + a) \sqrt{(n_i + 2a) \lambda \pi}} \\ &\quad \left\{ 1 + \frac{(\theta_i - \mu_i)^2}{(n_i + 2a) \delta^{*2} \lambda / (1 + \lambda)} \right\}^{-(n_i + 2a + 1)/2} \end{aligned}$$

which can be seen to be a scaled t-distribution with $n_i + 2a$ degrees of freedom and scale parameter $\delta^* \sqrt{\lambda / (1 + \lambda)}$ with $\delta^{*2} = \delta^2 / (n_i + 2a)$. Hence,

$$\begin{aligned} E(\sigma_i^{-1} | X_i, S_i, \mathbf{B}) &= \frac{\Gamma((n_i + 1)/2 + a) (\delta^2/2)^{-\{(n_i + 1)/2 + a\}}}{\Gamma(n_i/2 + a) (\delta^2/2)^{-(n_i/2 + a)}} \\ &= \frac{\Gamma((n_i + 1)/2 + a)}{\Gamma(n_i/2 + a)} \frac{2}{\delta^* \cdot n_i + 2a}. \end{aligned}$$

In this context, choosing

$$k = k(\mathbf{B}) = \left\{ 1 + \frac{t_{\alpha/2}^2}{n_i - 1} \right\}^{-(n_i + 2a + 1)/2} \sqrt{\frac{1 + \lambda}{\lambda} \frac{1}{\sqrt{2\pi}}},$$

the confidence interval in (8) simplifies to

$$C_i(\mathbf{B}) \equiv \left\{ \theta_i : \frac{|\theta_i - \mu_i|}{\frac{\lambda}{1 + \lambda} \frac{(n_i + 2a)\delta_i^{*2}}{n_i - 1}} \leq t_{\alpha/2} \right\}. \quad (\text{C.6})$$

Using the similar arguments as before and noting that $(n_i + 2a)\delta_i^{*2} \geq (n_i - 1)S_i^2$, we have $P\{C_i(\mathbf{B})\} \geq P(D_i) = 1 - \alpha$ where D_i is the confidence interval in (20). When \mathbf{B} is unknown, we replace \mathbf{B} by its marginal maximum likelihood estimate $\hat{\mathbf{B}}$. It is expected that the pooling technique will result in an error small enough so that $P\{C_i(\hat{\mathbf{B}})\} \approx P\{C_i(\mathbf{B})\} \geq 1 - \alpha$.

References

- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 95, 28-36.
- Bell, W. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. Technical Report U.S. Census Bureau.
- Casella, G., and Hwang, J. (1991). Evaluating confidence sets using loss functions. *Statistica Sinica*, 1, 159-173.
- Chatterjee, S., Lahiri, P. and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics*, 36, 1221-1245.
- Cho, M., Eltinge, J., Gershunskaya, J. and Huff, L. (2002). Evaluation of generalized variance function estimators for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 534-539.
- Fay, R., and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Gershunskaya, J., and Lahiri, P. (2005). Variance estimation for domains in the U.S. current employment statistics program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3044-3051.
- Ghosh, M., and Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 54-76.
- Hall, P., and Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested-error regression models. *Annals of Statistics*, 34, 1733-1750.
- Huff, L., Eltinge, J. and Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1519-1524.
- Hwang, J., Qiu, J. and Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both mean and variances. *Journal of the Royal Statistical Society, B*, 71, 265-285.
- Joshi, V. (1969). Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *The Annals of Mathematical Statistics*, 40, 1042-1067.
- Maples, J., Bell, W. and Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the American community survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 5056-5067.
- Otto, M., and Bell, W. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 160-165.
- Pfeffermann, D. (2002). Small area estimation - New developments and directions. *International Statistical Review*, 70, 125-143.
- Prasad, N., and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Qiu, J., and Hwang, J. (2007). Sharp simultaneous intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, 63, 767-776.
- Rao, J. (2003). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 145-169.
- Rivest, L.-P., and Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*.
- Robert, C., and Casella, G. (2004). *Monte Carlo Statistical Methods* (Second edition).
- Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82, 499-508.
- Wang, J., and Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *Journal of the American Statistical Association*, 98, 716-723.
- You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 1, 97-103.

Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data

Dan Liao and Richard Valliant¹

Abstract

Collinearities among explanatory variables in linear regression models affect estimates from survey data just as they do in non-survey data. Undesirable effects are unnecessarily inflated standard errors, spuriously low or high *t*-statistics, and parameter estimates with illogical signs. The available collinearity diagnostics are not generally appropriate for survey data because the variance estimators they incorporate do not properly account for stratification, clustering, and survey weights. In this article, we derive condition indexes and variance decompositions to diagnose collinearity problems in complex survey data. The adapted diagnostics are illustrated with data based on a survey of health characteristics.

Key Words: Diagnostics for survey data; Multicollinearity; Singular value decomposition; Variance inflation.

1. Introduction

When predictor variables in a regression model are correlated with each other, this condition is referred to as collinearity. Undesirable side effects of collinearity are unnecessarily high standard errors, spuriously low or high *t*-statistics, and parameter estimates with illogical signs or ones that are overly sensitive to small changes in data values. In experimental design, it may be possible to create situations where the explanatory variables are orthogonal to each other, but this is not true with observational data. Belsley (1991) noted that: "... in nonexperimental sciences, ..., collinearity is a natural law in the data set resulting from the uncontrollable operations of the data-generating mechanism and is simply a painful and unavoidable fact of life." In many surveys, variables that are substantially correlated are collected for analysis. Few analysts of survey data have escaped the problem of collinearity in regression estimation, and the presence of this problem encumbers precise statistical explanation of the relationships between predictors and responses.

Although many regression diagnostics have been developed for non-survey data, there are considerably fewer for survey data. The few articles that are available concentrate on identifying influential points and influential groups with abnormal data values or survey weights. Elliot (2007) developed Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. Li (2007a, b) and Li and Valliant (2009, 2011) extended a series of traditional diagnostic techniques to regression on complex survey data. Their papers cover residuals and leverages, several diagnostics based on case-deletion (DFBETA, DFBETAS, DFFIT, DFFITS, and Cook's Distance), and the forward search approach. Although an extensive literature in applied

statistics provides valuable suggestions and guidelines for data analysts to diagnose the presence of collinearity (e.g., Belsley, Kuh and Welsch 1980; Belsley 1991; Farrar and Glauber 1967; Fox 1986; Theil 1971), almost none of this research touches upon diagnostics for collinearity when fitting models with survey data. One prior, survey-related paper on collinearity problems is (Liao and Valliant 2012) which adapted variance inflation factors for linear models fitted with survey data.

Suppose the underlying structural model in the super-population is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. The matrix \mathbf{X} is an $n \times p$ matrix of predictors with n being the sample size; $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. The error terms in the model have a general variance structure $\mathbf{e} \sim (0, \sigma^2 \mathbf{R})$ where σ^2 is an unknown constant and \mathbf{R} is a unknown $n \times n$ covariance matrix. Define \mathbf{W} to be the diagonal matrix of survey weights. We assume throughout that the survey weights are constructed in such a way that they can be used for estimating finite population totals. The survey weighted least squares (SWLS) estimator is

$$\hat{\boldsymbol{\beta}}_{\text{sw}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \equiv \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y},$$

assuming $\mathbf{A} = \mathbf{X}^T \mathbf{W}^{-1} \mathbf{X}$ is invertible. Fuller (2002) describes the properties of this estimator. The estimator $\hat{\boldsymbol{\beta}}_{\text{sw}}$ is model unbiased for $\boldsymbol{\beta}$ under the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ regardless of whether $\text{Var}_M(\mathbf{e}) = \sigma^2 \mathbf{R}$ is specified correctly or not, and is approximately design-unbiased for the census parameter $\mathbf{B}_U = (\mathbf{X}_U^T \mathbf{X}_U)^{-1} \mathbf{X}_U^T \mathbf{Y}_U$, in the finite population U of N units. The finite population values of the response vector and matrix of predictors are $\mathbf{Y}_U = (Y_1, \dots, Y_N)^T$, and $\mathbf{X}_U = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ with \mathbf{X}_k being the $N \times 1$ vector of values for covariate k .

The remainder of the paper is organized as follows. Section 2 reviews results on condition numbers and variance

1. Dan Liao, RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005. E-mail: dliao@rti.org; Richard Valliant, University of Michigan and University of Maryland, Joint Program in Survey Methodology, 1218 Lefrak Hall, College Park, MD, 20742.

decompositions for ordinary least squares. These are extended to be appropriate for survey estimation in section 3. The fourth section gives some numerical illustrations of the techniques. Section 5 is a conclusion. In most derivations, we use model-based calculations since the forms of the model-variances are useful for understanding the effects of collinearity. However, when presenting variance decompositions, we use estimators that have both model- and design-based justifications.

2. Condition indexes and variance decompositions in ordinary least squares estimation

In this section we briefly review techniques for diagnosing collinearity in ordinary least squares (OLS) estimation based on condition indexes and variance decompositions. These methods will be extended in section 3 to cover complex survey data.

2.1 Eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$

When there is an exact (perfect) collinear relation in the $n \times p$ data matrix \mathbf{X} , we can find a set of values, $\mathbf{v} = (v_1, \dots, v_p)$, not all zero, such that

$$v_1\mathbf{X}_1 + \dots + v_p\mathbf{X}_p = \mathbf{0}, \quad \text{or} \quad \mathbf{X}\mathbf{v} = \mathbf{0}. \quad (1)$$

However, in practice, when there exists no exact collinearity but some near dependencies in the data matrix, it may be possible to find one or more non-zero vectors \mathbf{v} such that $\mathbf{X}\mathbf{v} = \mathbf{a}$ with $\mathbf{a} \neq \mathbf{0}$ but close to $\mathbf{0}$. Alternatively, we might say that a near dependency exists if the length of vector \mathbf{a} , $\|\mathbf{a}\|$, is small. To normalize the problem of finding the set of \mathbf{v} 's that makes $\|\mathbf{a}\|$ small, we consider only \mathbf{v} with unit length, that is, with $\|\mathbf{v}\| = 1$. Belsley (1991) discusses the connection of the eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$ with the normalized vector \mathbf{v} and $\|\mathbf{a}\|$. The minimum length $\|\mathbf{a}\|$ is simply the positive square root of the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}$. The \mathbf{v} that produces the \mathbf{a} with minimum length must be the eigenvector of $\mathbf{X}^T\mathbf{X}$ that corresponds to the smallest eigenvalue. As discussed in the next section, the eigenvalues and eigenvectors of \mathbf{X} are related to those of $\mathbf{X}^T\mathbf{X}$ and have some advantages when examining collinearity.

2.2 Singular-value decomposition, condition number and condition indexes

The singular-value decomposition (SVD) of matrix \mathbf{X} is very closely allied to the eigensystem of $\mathbf{X}^T\mathbf{X}$, but with its own advantages. The $n \times p$ matrix \mathbf{X} can be decomposed as $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and $\mathbf{D} = \text{diag}(\mu_1, \dots, \mu_p)$ is the diagonal matrix of singular values

(or eigenvalues) of \mathbf{X} . Here, the three components in the decomposition are matrices with very special, highly exploitable properties: \mathbf{U} is $n \times p$ (the same size as \mathbf{X}) and is column orthogonal; \mathbf{V} is $p \times p$ and both row and column orthogonal; \mathbf{D} is $p \times p$, nonnegative and diagonal. Belsley *et al.* (1980) felt that the SVD of \mathbf{X} has several advantages over the eigen system of $\mathbf{X}^T\mathbf{X}$, for the sake of both statistical usages and computational complexity. For prediction, \mathbf{X} is the focus not the cross-product matrix $\mathbf{X}^T\mathbf{X}$ since $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. In addition, the lengths $\|\mathbf{a}\|$ of the linear combinations (1) of \mathbf{X} that relate to collinearity are properly defined in terms of the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$, which are the singular values of \mathbf{X} . A secondary consideration, given current computing power, is that the singular value decomposition of \mathbf{X} avoids the additional computational burden of forming $\mathbf{X}^T\mathbf{X}$, an operation involving np^2 unneeded sums and products, which may lead to unnecessary truncation error.

The condition number of \mathbf{X} is defined as $\kappa(\mathbf{X}) = \mu_{\max} / \mu_{\min}$, where μ_{\max} and μ_{\min} are the maximum and minimum singular values of \mathbf{X} . Condition indexes are defined as $\eta_k = \mu_{\max} / \mu_k$. The closer that μ_{\min} is to zero, the nearer $\mathbf{X}^T\mathbf{X}$ is to being singular. Empirically, if a value of κ or η exceeds a cutoff value of, say, 10 to 30, two or more columns of \mathbf{X} have moderate or strong relations. The simultaneous occurrence of several large η_k 's is always remarkable for the existence of more than one near dependency.

One issue with the SVD is whether the \mathbf{X} 's should be centered around their means. Marquardt (1980) maintained that the centering of observations removes nonessential ill conditioning. In contrast, Belsley (1984) argues that mean-centering typically masks the role of the constant term in any underlying near-dependencies. A typical case is a regression with dummy variables. For example, if gender is one of the independent variables in a regression and most of the cases are male (or female), then the dummy for gender can be strongly collinear with the intercept. The discussions following Belsley (1984) illustrate the differences of opinion that occur among practitioners (Wood 1984; Snee and Marquardt 1984; Cook 1984). Moreover, in linear regression analysis, Wissmann, Toutenburg and Shalabh (2007) found that the degree of multicollinearity with dummy variables may be influenced by the choice of reference category. In this article, we do not center the \mathbf{X} 's but will illustrate the effect of the choice of reference category in section 4.

Another problem with the condition number is that it is affected by the scale of the x measurements (Steward 1987). By scaling down any column of \mathbf{X} , the condition number can be made arbitrarily large. This situation is known as *artificial ill-conditioning*. Belsley (1991) suggests

scaling each column of the design matrix \mathbf{X} using the Euclidean norm of each column before computing the condition number. This method is implemented in SAS and the package *perturb* of the statistical software R (Hendrickx 2010). Both use the root mean square of each column for scaling as its standard procedure. The condition number and condition indexes of the scaled matrix \mathbf{X} are referred to as the *scaled condition number* and *scaled condition indexes* of the matrix \mathbf{X} . Similarly, the variance decomposition proportions relevant to the scaled \mathbf{X} (which will be discussed in next section) will be called the *scaled variance decomposition proportions*.

2.3 Variance decomposition method

To assess the extent to which near dependencies (*i.e.*, having high condition indexes of \mathbf{X} and $\mathbf{X}^T\mathbf{X}$) degrade the estimated variance of each regression coefficient, Belsley *et al.* (1980) reinterpreted and extended the work of Silvey (1969) by decomposing a coefficient variance into a sum of terms each of which is associated with a singular value. In the remainder of this section, we review the results of ordinary least squares (OLS) under the model $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{Var}_M(\mathbf{Y}) = \sigma^2\mathbf{I}_n$ where \mathbf{I}_n is the $n \times n$ identity matrix. These results will be extended to survey weighted least squares in section 3. Recall that the model variance-covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is $\text{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Using the SVD, $\mathbf{X} = \mathbf{UDV}^T$, $\text{Var}_M(\hat{\boldsymbol{\beta}})$ can be written as:

$$\text{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2[(\mathbf{UDV}^T)^T(\mathbf{UDV}^T)]^{-1} = \sigma^2\mathbf{VD}^{-2}\mathbf{V}^T \quad (2)$$

and the k^{th} diagonal element in $\text{Var}_M(\hat{\boldsymbol{\beta}})$ is the estimated variance for the k^{th} coefficient, $\hat{\beta}_k$. Using (2), $\text{Var}_M(\hat{\beta}_k)$ can be expressed as:

$$\text{Var}_M(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^p \frac{v_{kj}^2}{\mu_j^2} \quad (3)$$

where $\mathbf{V} = (v_{kj})_{p \times p}$. Let $\phi_{kj} = v_{kj}^2 / \mu_j^2$, $\phi_k = \sum_{j=1}^p \phi_{kj}$ and $\mathbf{Q} = (\phi_{kj})_{p \times p} = (\mathbf{VD}^{-1}) \cdot (\mathbf{VD}^{-1})$, where \cdot is the Hadamard (elementwise) product. The variance-decomposition proportions are $\pi_{jk} = \phi_{jk} / \phi_k$, which is the proportion of the variance of the k^{th} regression coefficient associated with the j^{th} component of its decomposition in (3). Denote the *variance decomposition proportion matrix* as $\boldsymbol{\Pi} = (\pi_{jk})_{p \times p} = \mathbf{Q}^T \bar{\mathbf{Q}}^{-1}$, where $\bar{\mathbf{Q}}$ is the diagonal matrix with the row sums of \mathbf{Q} on the main diagonal and 0 elsewhere.

If the model is $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, $\text{Var}_M(\mathbf{Y}) = \sigma^2\mathbf{W}^{-1}$ and weighted least squares is used, then $\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$ and $\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{WLS}}) = \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$. The decomposition in (3) holds with $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$ being decomposed

as $\tilde{\mathbf{X}} = \mathbf{UDV}^T$. However, in survey applications, it will virtually never be the case that the covariance matrix of \mathbf{Y} is $\sigma^2\mathbf{W}^{-1}$ if \mathbf{W} is the matrix of survey weights. Section 3 covers the more realistic case.

In the variance decomposition (3), other things being equal, a small singular value μ_j can lead to a large component of $\text{Var}(\hat{\beta}_k)$. However, if v_{kj} is small too, then $\text{Var}(\hat{\beta}_k)$ may not be affected by a small μ_j . One extreme case is when $v_{kj} = 0$. Suppose the k^{th} and j^{th} columns of \mathbf{X} belong to separate orthogonal blocks. Let $\mathbf{X} \equiv [\mathbf{X}_1, \mathbf{X}_2]$ with $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$ and let the singular-value decompositions of \mathbf{X}_1 and \mathbf{X}_2 be given, respectively, as $\mathbf{X}_1 = \mathbf{U}_1\mathbf{D}_{11}\mathbf{V}_{11}^T$ and $\mathbf{X}_2 = \mathbf{U}_2\mathbf{D}_{22}\mathbf{V}_{22}^T$. Since \mathbf{U}_1 and \mathbf{U}_2 are the orthogonal bases for the space spanned by the columns of \mathbf{X}_1 and \mathbf{X}_2 respectively, $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$ implies $\mathbf{U}_1^T\mathbf{U}_2 = \mathbf{0}$ and $\mathbf{U} \equiv [\mathbf{U}_1, \mathbf{U}_2]$ is column orthogonal. The singular value decomposition of \mathbf{X} is simply $\mathbf{X} = \mathbf{UDU}_2^T$, with:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{22} \end{bmatrix} \quad (4)$$

and

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{22} \end{bmatrix}. \quad (5)$$

Thus $\mathbf{V}_{12} = \mathbf{0}$. An analogous result clearly applies to any number of mutually orthogonal subgroups. Hence, if all the columns in \mathbf{X} are orthogonal, all the $v_{kj} = 0$ when $k \neq j$ and $\pi_{kj} = 0$ likewise. When v_{kj} is nonzero, this is a signal that predictors k and j are not orthogonal.

Since at least one v_{kj} must be nonzero in (3), this implies that a high proportion of any variance can be associated with a large singular value even when there is no collinearity. The standard approach is to check a high condition index associated with a large proportion of the variance of two or more coefficients when diagnosing collinearity, since there must be two or more columns of \mathbf{X} involved to make a near dependency. Belsley *et al.* (1980) suggested showing the matrix $\boldsymbol{\Pi}$ and condition indexes of \mathbf{X} in a variance decomposition table as below. If two or more elements in the j^{th} row of matrix $\boldsymbol{\Pi}$ are relatively large and its associated condition index η_j is large too, it signals that near dependencies are influencing regression estimates.

Condition Index	Proportions of variance			
	$\text{Var}_M(\hat{\beta}_1)$	$\text{Var}_M(\hat{\beta}_2)$...	$\text{Var}_M(\hat{\beta}_p)$
η_1	π_{11}	π_{12}	...	π_{1p}
η_2	π_{21}	π_{22}	...	π_{2p}
\vdots	\vdots	\vdots		\vdots
η_p	π_{p1}	π_{p2}	...	π_{pp}

3. Adaptation in survey-weighted least squares

3.1 Condition indexes and variance decomposition proportions

In survey-weighted least squares (SWLS), we are more interested in the collinear relations among the columns in the matrix $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$ instead of \mathbf{X} , since $\hat{\beta}_{\text{SW}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{Y}}$. Define the singular value decomposition of $\tilde{\mathbf{X}}$ to be $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where \mathbf{U} , \mathbf{V} , and \mathbf{D} are usually different from the ones of \mathbf{X} , due to the unequal survey weights.

The condition number of $\tilde{\mathbf{X}}$ is defined as $\kappa(\tilde{\mathbf{X}}) = \mu_{\max} / \mu_{\min}$, where μ_{\max} and μ_{\min} are maximum and minimum singular values of $\tilde{\mathbf{X}}$. The condition number of $\tilde{\mathbf{X}}$ is also usually different from the condition number of the data matrix \mathbf{X} due to unequal survey weights. Condition indexes are defined as

$$\eta_k = \mu_{\max} / \mu_k, \quad k = 1, \dots, p \quad (6)$$

where μ_k is one of the singular values of $\tilde{\mathbf{X}}$. The scaled condition indexes and condition numbers are the condition indexes and condition numbers of the scaled $\tilde{\mathbf{X}}$.

Based on the extrema of the ratio of quadratic forms (Lin 1984), the condition number $\kappa(\tilde{\mathbf{X}})$ is bounded in the range of:

$$\frac{w_{\min}^{1/2}}{w_{\max}^{1/2}} \kappa(\mathbf{X}) \leq \kappa(\tilde{\mathbf{X}}) \leq \frac{w_{\max}^{1/2}}{w_{\min}^{1/2}} \kappa(\mathbf{X}), \quad (7)$$

where w_{\min} and w_{\max} are the minimum and maximum survey weights. This expression indicates that if the survey weights do not vary too much, the condition number in SWLS resembles the one in OLS. However, in a sample with a wide range of survey weights, the condition number can be very different between SWLS and OLS. When SWLS has a large condition number, OLS might not. In the case of exact linear dependence among the columns of \mathbf{X} , the columns of $\tilde{\mathbf{X}}$ will also be linearly dependent. In this extreme case at least one eigenvalue of \mathbf{X} will be zero, and both $\kappa(\mathbf{X})$ and $\kappa(\tilde{\mathbf{X}})$ will be infinite. As in OLS, large values of κ or of the η_k 's of 10 or more may signal that two or more columns of \mathbf{X} have moderate to strong dependencies.

The model variance of the SWLS parameter estimator under a model with $\text{Var}_M(\mathbf{e}) = \sigma^2 \mathbf{R}$ is:

$$\begin{aligned} \text{Var}_M(\hat{\beta}_{\text{SW}}) &= \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{R} \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \\ &= \sigma^2 (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{G}, \end{aligned} \quad (8)$$

where

$$\mathbf{G} = (\mathbf{g}_{ij})_{p \times p} = \mathbf{X}^T \mathbf{W} \mathbf{R} \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \quad (9)$$

is the *misspecification effect* (MEFF) that represents the inflation factor needed to correct standard results for the effect of intracluster correlation in clustered survey data and for the fact that $\text{Var}_M(\mathbf{e}) = \sigma^2 \mathbf{R}$ and not $\sigma^2 \mathbf{W}^{-1}$ (Scott and Holt 1982).

Using the SVD of $\tilde{\mathbf{X}}$, we can rewrite $\text{Var}_M(\hat{\beta}_{\text{SW}})$ as

$$\text{Var}_M(\hat{\beta}_{\text{SW}}) = \sigma^2 \mathbf{V} \mathbf{D}^{-2} \mathbf{V}^T \mathbf{G}. \quad (10)$$

The k^{th} diagonal element in $\text{Var}_M(\hat{\beta}_{\text{SW}})$ is the estimated variance for the k^{th} coefficient, $\hat{\beta}_k$. Using (10), $\text{Var}_M(\hat{\beta}_k)$ can be expressed as:

$$\text{Var}_M(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^p \frac{v_{kj}}{\mu_j^2} \lambda_{kj} \quad (11)$$

where $\lambda_{kj} = \sum_{i=1}^p v_{ij} g_{ik}$. If $\mathbf{R} = \mathbf{W}^{-1}$, then $\mathbf{G} = \mathbf{I}_p$, $\lambda_{kj} = v_{kj}$, and (11) reduces to (3). However, the situation is more complicated when \mathbf{G} is not the identity matrix, *i.e.*, when the complex design affects the variance of an estimated regression coefficient. If predictors k and j are orthogonal, $v_{kj} = 0$ for $k \neq j$ and the variance in (11) depends only on the k^{th} singular value and is unaffected by g_{ij} 's that are non-zero. If predictor k and several j 's are not orthogonal, then λ_{kj} has contributions from all of those eigenvectors and from the off-diagonal elements of the MEFF matrix \mathbf{G} . The term λ_{kj} then measures both non-orthogonality of x 's and effects of the complex design.

Consequently, we can define variance decomposition proportions analogous to those for OLS but their interpretation is less straightforward. Let $\phi_{kj} = v_{kj} \lambda_{kj} / \mu_j^2$, $\Phi_k = \sum_{j=1}^p \phi_{kj}$ and $\mathbf{Q} = (\Phi_k)_{p \times p} = (\mathbf{V} \mathbf{D}^{-2}) \cdot (\mathbf{V}^T \mathbf{G})^T$. The variance-decomposition proportions are $\pi_{jk} = \phi_{jk} / \Phi_k$, which is the proportion of the variance of the k^{th} regression coefficient associated with the j^{th} component of its decomposition in (11). Denote the variance decomposition proportion matrix as

$$\Pi = (\pi_{jk})_{p \times p} = \mathbf{Q}^T \bar{\mathbf{Q}}^{-1}, \quad (12)$$

where $\bar{\mathbf{Q}}$ is the diagonal matrix with the row sums of \mathbf{Q} on the main diagonal and 0 elsewhere. The interpretation of the proportions in (12) is not as clear-cut as for OLS because the effect of the MEFF matrix. Section 3.2 discusses the interpretation in more detail in the context of stratified cluster sampling.

Analogous to the method for OLS regression, a variance decomposition table can be formed like the one at the end of section 2. When two or more independent variables are collinear (or "nearly dependent"), one singular value should make a large contribution to the variance of the parameter estimates associated with those variables. For example, if the proportions π_{31} and π_{32} for the variances of $\hat{\beta}_{\text{SW}1}$ and

$\hat{\beta}_{\text{SW}2}$ are large, this would say that the third singular value makes a large contribution to both variances and that the first and second predictors in the regression are, to some extent, collinear. As shown in section 2.3, when the k^{th} and j^{th} columns in \mathbf{X} are orthogonal, $v_{kj} = 0$ and the j^{th} singular value's decomposition proportion π_{jk} on $\text{Var}(\hat{\beta}_k)$ will be 0.

Several special cases are worth noting. If $\mathbf{R} = \mathbf{W}^{-1}$ as assumed in WLS, then $\mathbf{G} = \mathbf{I}$. The variance decomposition in (11) has the same form as (2) in OLS. However, having $\mathbf{R} = \mathbf{W}^{-1}$ in survey data would be unusual since survey weights are not typically computed based on the variance structure of a model. Note that \mathbf{V} is still different from the one in OLS and is one component of the SVD of $\tilde{\mathbf{X}}$ instead of \mathbf{X} . Another special case here is when $\mathbf{R} = \mathbf{I}$ and the survey weights are equal, in which case the OLS results can be used. However, when the survey weights are unequal, even when $\mathbf{R} = \mathbf{I}$, the variance decomposition in (11) is different from (2) in OLS since $\mathbf{G} \neq \mathbf{I}$. In the next section, we will consider some special models that take the population features such as clusters and strata into account when estimating this variance decomposition.

3.2 Variance decomposition for a model with stratified clustering

The model variance of $\hat{\beta}_{\text{SW}}$ in (8) contains the unknown \mathbf{R} that must be estimated. In this section, we present an estimator for $\hat{\beta}_{\text{SW}}$ that is appropriate for a model with stratified clustering. The variance estimator has both model-based and design-based justification. Suppose that in a stratified multistage sampling design, there are strata $h = 1, \dots, H$ in the population, clusters $i = 1, \dots, N_h$ in stratum h and units $t = 1, \dots, M_{hi}$ in cluster hi . We select clusters $i = 1, \dots, n_h$ in stratum h and units $t = 1, \dots, m_{hi}$ in cluster hi . Denote the set of sample clusters in stratum h by s_h and the sample of units in cluster hi as s_{hi} . The total number of sample units in stratum h is $m_h = \sum_{i \in s_h} m_{hi}$, and the total in the sample is $m = \sum_{h=1}^H m_h$. Assume that clusters are selected with varying probabilities and with replacement within strata and independently between strata. The model we consider is:

$$\begin{aligned} E_M(Y_{hit}) &= \mathbf{x}_{hit}^T \boldsymbol{\beta} \\ h &= 1, \dots, H, i = 1, \dots, N_h, t = 1, \dots, M_{hi} \\ \text{Cov}_M(\epsilon_{hit}, \epsilon_{h'i't'}) &= 0 \\ \text{where } \epsilon_{hit} &= Y_{hit} - \mathbf{x}_{hit}^T \boldsymbol{\beta}, \quad i \neq i' \\ \text{Cov}_M(\epsilon_{hit}, \epsilon_{h'i't'}) &= 0 \quad h \neq h'. \end{aligned} \quad (13)$$

Units within each cluster are assumed to be correlated but the particular form of the covariances does not have to be

specified for this analysis. The estimator $\hat{\beta}_{\text{SW}}$ of the regression parameter can be written as:

$$\hat{\beta}_{\text{SW}} = \sum_{h=1}^H \sum_{i \in s_h} (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{Y}_{hi} \quad (14)$$

where \mathbf{X}_{hi} is the $m_{hi} \times p$ matrix of covariates for sample units in cluster hi , $\mathbf{W}_{hi} = \text{diag}(w_t)$, $t \in s_{hi}$, is the diagonal matrix of survey weights for units in cluster hi and \mathbf{Y}_{hi} is the $m_{hi} \times 1$ vector of response variables in cluster hi . The model variance of $\hat{\beta}_{\text{SW}}$ is:

$$\text{Var}_M(\hat{\beta}_{\text{SW}}) = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{G}_{st} \quad (15)$$

where

$$\begin{aligned} \mathbf{G}_{st} &= \left[\sum_{h=1}^H \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{R}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi} \right] (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \\ &= \left[\sum_{h=1}^H \mathbf{X}_h^T \mathbf{W}_h \mathbf{R}_h \mathbf{W}_h \mathbf{X}_h \right] (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \end{aligned} \quad (16)$$

with $\mathbf{R}_{hi} = \text{Var}_M(\mathbf{Y}_{hi})$, $\mathbf{W}_h = \text{diag}(\mathbf{W}_{hi})$, and $\mathbf{R}_h = \text{Blkdiag}(\mathbf{R}_{hi})$, $\mathbf{W}_h = \text{diag}(\mathbf{W}_{hi})$, $\mathbf{X}_h^T = (\mathbf{X}_{h1}^T, \mathbf{X}_{h2}^T, \dots, \mathbf{X}_{hn_h}^T)$, $i \in s_h$. Expression (16) is a special case of (9) with $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T, \dots, \mathbf{X}_H^T)$, where \mathbf{X}_h is the $m_h \times p$ matrix of covariates for sample units in stratum h , $\mathbf{W} = \text{diag}(\mathbf{W}_{hi})$, for $h = 1, \dots, H$ and $i \in s_h$ and $\mathbf{R} = \text{Blkdiag}(\mathbf{R}_{hi})$.

Based on the development in Scott and Holt (1982, section 4), the MEFF matrix \mathbf{G}_{st} can be rewritten for a special case of \mathbf{R}_h in a way that will make the decomposition proportions in (12) more understandable. Consider the special case of (13) with

$$\text{Cov}_M(\epsilon_{hi}) = \sigma^2(1 - \rho) \mathbf{I}_{m_{hi}} + \sigma^2 \rho \mathbf{I}_{m_{hi}} \mathbf{1}_{m_{hi}}^T$$

where $\mathbf{I}_{m_{hi}}$ is the $m_{hi} \times m_{hi}$ identity matrix and $\mathbf{1}_{m_{hi}}$ is a vector of m_{hi} 1's. In that case,

$$\begin{aligned} \mathbf{X}_h^T \mathbf{W}_h \mathbf{R}_h \mathbf{W}_h \mathbf{X}_h &= (1 - \rho) \mathbf{X}_h^T \mathbf{W}_h^2 \mathbf{X}_h \\ &\quad + \rho \sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{W}_{Bhi}^2 \mathbf{X}_{Bhi} \end{aligned}$$

where $\mathbf{X}_{Bhi} = m_{hi}^{-1} \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T \mathbf{X}_{hi}$. Suppose that the sample is self-weighting so that $\mathbf{W}_{hi} = w \mathbf{I}_{m_{hi}}$. After some simplification, it follows that

$$\mathbf{G}_{st} = w[\mathbf{I}_p + (\mathbf{M} - \mathbf{I}_p) \rho]$$

where \mathbf{I}_p is the $p \times p$ identity matrix and $\mathbf{M} = (\sum_{h=1}^H \sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{X}_{Bhi}) (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$. Thus, if the sample is self-weighting and ρ is very small, then $\mathbf{G}_{st} \approx w \mathbf{I}_p$ and

$\text{Var}_M(\hat{\beta}_{\text{SW}})$ in (15) will be approximately the same as the OLS variance. If so, the SWLS variance decomposition proportions will be similar to the OLS proportions. In regression problems, ρ often is small since it is the correlation of the errors, $\varepsilon_{hi} = Y_{hi} - \mathbf{x}_{hi}^T \beta$, for different units rather than for \mathbf{Y}_{hi} 's. This is related to the phenomenon that design effects for regression coefficients are often smaller than for means—a fact first noted by Kish and Frankel (1974). In applications where ρ is larger, the variance decomposition proportions in (12) will still be useful in identifying collinearity although they will be affected by departures of the model errors from independence.

Denote the cluster-level residuals as a vector, $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\beta}_{\text{SW}}$. The estimator of (15) that we consider was originally derived from design-based considerations. A linearization estimator, appropriate when clusters are selected with replacement, is:

$$\text{var}_L(\hat{\beta}_{\text{SW}}) = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \hat{\mathbf{G}}_L \quad (17)$$

with the estimated misspecification effect as

$$\hat{\mathbf{G}}_L = (\hat{g}_{ij})_{p \times p} = \left[\sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{i \in s_h} (\mathbf{z}_{hi}^* - \bar{\mathbf{z}}_h^*)(\mathbf{z}_{hi}^* - \bar{\mathbf{z}}_h^*)^T \right] (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}, \quad (18)$$

where $\bar{\mathbf{z}}_h^* = 1/n_h \sum_{i \in s_h} \mathbf{z}_{hi}^*$ and $\mathbf{z}_{hi}^* = \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{e}_{hi}$ with $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\beta}_{\text{SW}}$, and the variance-covariance matrix \mathbf{R} can be estimated by

$$\hat{\mathbf{R}} = \sum_{h=1}^H \frac{n_h}{n_h - 1} \left[\text{Blkdiag}(\mathbf{e}_{hi} \mathbf{e}_{hi}^T) - \frac{1}{n_h} \mathbf{e}_h \mathbf{e}_h^T \right].$$

Expression (17) is used by the Stata and SUDAAN packages, among others. The estimator $\text{var}_L(\hat{\beta}_{\text{SW}})$ is consistent and approximately design-unbiased under a design where clusters are selected with replacement (Fuller 2002). The estimator in (17) is also an approximately model-unbiased estimator of (15) (see Liao 2010). Since the estimator $\text{var}_L(\hat{\beta}_{\text{SW}})$ is also currently available in software packages, we will use it in the empirical work in section 4.

Using (12) derived in section 2, the variance decomposition proportion matrix Π for $\text{var}_L(\hat{\beta}_{\text{SW}})$ can then be written as

$$\Pi = (\pi_{jk})_{p \times p} = \mathbf{Q}_L^T \bar{\mathbf{Q}}_L^{-1} \quad (19)$$

with $\mathbf{Q}_L = (\phi_{kj})_{p \times p} = (\mathbf{V} \mathbf{D}^{-2}) \cdot (\mathbf{V}^T \hat{\mathbf{G}}_L)^T$ and $\bar{\mathbf{Q}}_L$ is the diagonal matrix with the row sums of \mathbf{Q}_L on the main diagonal and 0 elsewhere.

4. Numerical illustrations

In this section, we will illustrate the collinearity measures described in section 3 and investigate their behaviors using the dietary intake data from 2007-2008 National Health and Nutrition Examination Survey (NHANES).

4.1 Description of the data

The dietary intake data are used to estimate the types and amounts of foods and beverages consumed during the 24-hour period prior to the interview (midnight to midnight), and to estimate intakes of energy, nutrients, and other food components from those foods and beverages. NHANES uses a complex, multistage, probability sampling design; oversampling of certain population subgroups is done to increase the reliability and precision of health status indicator estimates for these groups. Among the respondents who received the in-person interview in the mobile examination center (MEC), around 94% provided complete dietary intakes. The survey weights were constructed by taking MEC sample weights and further adjusting for the additional nonresponse and the differential allocation by day of the week for the dietary intake data collection. These weights are more variable than the MEC weights. The data set used in our study is a subset of 2007-2008 data composed of female respondents aged 26 to 40. Observations with missing values in the selected variables are excluded from the sample which finally contains 672 complete respondents. The final weights in our sample range from 6,028 to 330,067, with a ratio of 55:1. The U.S. National Center for Health Statistics recommends that the design of the sample is approximated by the stratified selection with replacement of 32 PSUs from 16 strata, with 2 PSUs within each stratum.

4.2 Study one: Correlated covariates

In the first empirical study, a linear regression model of respondent's body mass index (BMI) was considered. The explanatory variables considered included two demographic variables, respondent's age and race (Black/Non-black), four dummy variables for whether the respondent is on a special diet of any kind, on a low-calorie diet, on a low-fat diet, and on a low-carbohydrate diet (when he/she is on diet, value equals 1, otherwise 0), and ten daily total nutrition intake variables, consisting of total calories (100kcal), protein (100gm), carbohydrate (100gm), sugar (100gm), dietary fiber (100gm), alcohol (100gm), total fat (100gm), total saturated fatty acids (100gm), total monounsaturated fatty acids (100gm), and total polyunsaturated fatty acids (100gm). The correlation coefficients among these variables are displayed in Table 2. Note that the correlations among the daily total nutrition intake variables are often high. For

example, the correlations of the total fat intakes with total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids are 0.85, 0.97 and 0.93.

Three types of regressions were fitted for the selected sample to demonstrate different diagnostics. More details about these three regression types and their diagnostic statistics are displayed in Table 1.

TYPE1: OLS regression with estimated σ^2 ; the diagnostic statistics are obtained using the standard methods reviewed in section 2;

TYPE2: WLS regression with estimated σ^2 and assuming $\mathbf{R} = \mathbf{W}^{-1}$; the scaled condition indexes are estimated using (6) and the scaled variance decomposition proportions are estimated using (12). With $\mathbf{R} = \mathbf{W}^{-1}$, these are the variance decompositions that will be produced by standard software using WLS and specifying the weights to be the survey weights;

TYPE3: SWLS with estimated $\hat{\mathbf{R}}$; the scaled condition indexes are estimated using (6); the scaled variance decomposition proportions are estimated using (12).

Their diagnostic statistics, including the scaled condition indexes and variance decomposition proportions are reported in Tables 3, 4 and 5, respectively. To make the

tables more readable, only the proportions that are larger than 0.3 are shown. Proportions that are less than 0.3 are shown as dots. Note that some terms in decomposition (12) can be negative. This leads to the possibility of some “proportions” being greater than 1. This occurs in five cases in Table 5. Belsley *et al.* (1980) suggest that a condition index of 10 signals that collinearity has a moderate effect on standard errors; an index of 100 would indicate a serious effect. In this study, we consider a scaled condition index greater than 10 to be relatively large, and ones greater than 30 as large and remarkable. Furthermore, the large scaled variance-decomposition proportions (greater than 0.3) associated with each large scaled condition index will be used to identify those variates that are involved in a near dependency. The intracluster correlation of the residuals is shown in the last row of Table 6 under the column labeled “Original Model”. In the model used for Tables 3-5, $\rho = 0.0366$ as estimated from a model with random effects for clusters. As noted in section 3.2, when ρ is small and the sample is self-weighting, the SWLS decomposition proportions can be interpreted in the same way as those of OLS. Although the NHANES sample does not have equal weights, ρ is small in this example and the decomposition proportions should still provide useful information.

Table 1
Regression models and their collinearity diagnostic statistics used in this experimental study

Type	Regression Method	Weight matrix \mathbf{W}^a	$\text{var}(\hat{\beta})$	$\text{var}(\hat{\beta}_k)$	Matrix for Condition Indexes ^b	Variance Decomposition Proportion π_{jk}
TYPE1	OLS	\mathbf{I}	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$	$\hat{\sigma}^2\sum_{j=1}^p\frac{u_{2kj}^2}{\mu_j^2}$ ^c	$\mathbf{X}^T\mathbf{X}$	$\frac{u_{2kj}^2}{\mu_j^2} / \sum_{j=1}^p\frac{u_{2kj}^2}{\mu_j^2}$
TYPE2	WLS	\mathbf{W}	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\hat{\sigma}^2\sum_{j=1}^p\frac{u_{2kj}^2}{\mu_j^2}$ ^d	$\mathbf{X}^T\mathbf{W}\mathbf{X}$	$\frac{u_{2kj}^2}{\mu_j^2} / \sum_{j=1}^p\frac{u_{2kj}^2}{\mu_j^2}$
TYPE3	SWLS	\mathbf{W}	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{R}}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\hat{\sigma}^2\sum_{j=1}^p\frac{u_{2kj}\sum_{i=1}^p\hat{g}_{ik}u_{2ij}}{\mu_j^2}$ ^e	$\mathbf{X}^T\mathbf{W}\mathbf{X}$	$\frac{u_{2kj}\sum_{i=1}^p\hat{g}_{ik}u_{2ij}}{\mu_j^2} / \sum_{j=1}^p\frac{u_{2kj}\sum_{i=1}^p\hat{g}_{ik}u_{2ij}}{\mu_j^2}$

$$\hat{\mathbf{R}} = \sum_{h=1}^H \frac{n_h}{n_h - 1} \left[\text{Blkdiag}(\mathbf{e}_h\mathbf{e}_h^T) - \frac{1}{n_h}\mathbf{e}_h\mathbf{e}_h^T \right]$$

^a In all the regression models, the parameters are estimated by: $\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$.
^b The eigenvalues of this matrix will be used to compute the Condition Indexes for the corresponding regression model.
^c The terms u_{2kj} and μ_j are from the singular value decomposition of the data matrix \mathbf{X} .
^d The terms u_{2kj} and μ_j are from the singular value decomposition of the weighted data matrix $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$.
^e The terms u_{2kj} and μ_j are from the singular value decomposition (SVD) of the weighted data matrix $\tilde{\mathbf{X}}$. The term \hat{g}_{ik} is the unit element of misspecification effect matrix $\hat{\mathbf{G}}$.

In Tables 3, 4 and 5, the weighted regression methods, WLS and SWLS, used the survey-weighted data matrix $\tilde{\mathbf{X}}$ to obtain the condition indexes while the unweighted regression method, OLS, used the data matrix \mathbf{X} . The largest scaled condition index in WLS and SWLS is 566, which is slightly smaller than the one in OLS, 581. Both of these values are much larger than 30 and, thus, signal a severe near-dependency among the predictors in all three regression models. Such large condition numbers imply that the inverse of the design matrix, $\mathbf{X}^T\mathbf{W}\mathbf{X}$, may be numerically unstable, *i.e.*, small changes in the x data could make large changes in the elements of the inverse.

The values of the decomposition proportions for OLS and WLS are very similar and lead to the same predictors being identified as potentially collinear. Results for SWLS are somewhat different as sketched below. In OLS and WLS, six daily total nutrition intake variables—calorie, protein, carbohydrate, alcohol, dietary fiber and total fat—are involved in the dominant near-dependency that is associated with the largest scaled condition index. Four daily fat intake variables, total fat, total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids, are involved in the secondary near-dependency that is associated with the second largest scaled condition index. A moderate near-dependency between intercept and age is also shown in all three tables. The associated scaled condition index is equal to 38 in OLS and 37 in WLS and SWLS. However, when SWLS is used, sugar, total saturated fatty acids and total polyunsaturated fatty acids also appear to be involved in the dominant near-dependency as shown in Table 5. While, only three daily fat intake variables, total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids, are involved in the secondary near-dependency that is associated with the second largest scaled condition index. Thus, when OLS or WLS is used, the impact of near-dependency among sugar, total saturated fatty acids, total polyunsaturated fatty acids and the six daily total nutrition intake variables is not as strong as the ones in SWLS. If conventional OLS or WLS diagnostics are used for SWLS, this near-dependency might be overlooked.

Rather than using the scaled condition indexes and variance decomposition method (in Tables 3, 4 and 5), an analyst might attempt to identify collinearities by examining the unweighted correlation coefficient matrix in Table 2. Although the correlation coefficient matrix shows that almost all the daily total nutrition intake variables are highly or moderately pairwise correlated, it cannot be used to

reliably identify the near-dependencies among these variables when used in a regression. For example, the correlation coefficient between “on any diet” and “on low-calorie diet” is relatively large (0.73). This near dependency is associated with a scaled condition index equal to 11 (larger than 10, but less than the cutoff of 30) in OLS and WLS (shown in Table 3 and 4) and is associated with a scaled condition index equal to 2 (less than 10) in SWLS (shown in Table 5). The impact of this near dependency appears to be not very harmful no matter which regression method is used. On the other hand, alcohol is weakly correlated with all the daily total nutrition intake variables but is highly involved in the dominant near-dependency shown in the last row of Tables 3-5.

After the collinearity patterns are diagnosed, the common corrective action would be to drop the correlated variables, refit the model and reexamine standard errors, collinearity measures and other diagnostics. Omitting \mathbf{X} 's one at a time may be advisable because of the potentially complex interplay of explanatory variables. In this example, if the total fat intake is one of the key variables that an analyst feels must be kept, sugar might be dropped first followed by protein, calorie, alcohol, carbohydrate, total fat, dietary fiber, total monounsaturated fatty acids, total polyunsaturated fatty acids and total saturated fatty acids. Other remedies for collinearity could be to transform the data or use some specialized techniques such as ridge regression and mixed Bayesian modeling, which require extra (prior) information beyond the scope of most research and evaluations.

To demonstrate how the collinearity diagnostics can improve the regression results in this example, Table 6 presents the SWLS regression analysis output of the original models with all the explanatory variables and a reduced model with fewer explanatory variables. In the reduced model, all of the dietary intake variables are eliminated except total fat intake. After the number of correlated offending variables is reduced, the standard error of total fat intake is only the one forty-sixth of its standard error in the original model. The total fat intake becomes significant in the reduced model. The reduction of correlated variables appears to have substantially improved the accuracy of estimating the impact of total fat intake on BMI. Note that the collinearity diagnostics do not provide a unique path toward a final model. Different analysts may make different choices about whether particular predictors should be dropped or retained.

Table 2
Correlation coefficient matrix of the data matrix X

	age	black	on any diet	on low- calorie diet	on low- fat diet	on low- carb diet ^a	calorie	protein	Carbo- hydrate	sugar	fiber	alcohol	total. fat	sat. fat	mono. fat	poly. fat
age	1															
black	<i>b</i>	1														
on any diet	.	.	1													
on low-calorie diet	.	.	<i>0.87^c</i>	1												
on low-fat diet	1											
on low-carb diet	1										
calorie	1									
protein	<i>0.75</i>	1								
carb	<i>0.84</i>	<i>0.45</i>	1							
sugar	<i>0.58</i>	.	<i>0.84</i>	1						
fiber	<i>0.57</i>	<i>0.52</i>	<i>0.54</i>	.	1					
alcohol	1				
total.fat	<i>0.86</i>	<i>0.72</i>	<i>0.54</i>	.	<i>0.48</i>	.	1			
sat.fat ^d	<i>0.74</i>	<i>0.56</i>	<i>0.47</i>	.	<i>0.46</i>	.	<i>0.85</i>	1		
mono.fat ^e	<i>0.83</i>	<i>0.68</i>	<i>0.51</i>	.	<i>0.46</i>	.	<i>0.97</i>	<i>0.82</i>	1	
poly.fat ^f	<i>0.81</i>	<i>0.71</i>	<i>0.51</i>	.	<i>0.43</i>	.	<i>0.93</i>	<i>0.63</i>	<i>0.87</i>	1

^a The term "carb" stands for carbohydrate.

^b Correlation coefficients less than 0.3 are omitted in this table.

^c Correlation coefficients larger than 0.3 are italicized in this table.

^d Total Saturated Fatty Acids.

^e Total Monounsaturated Fatty Acids.

^f Total Polyunsaturated Fatty Acids.

Table 3
Scaled condition indexes and variance decomposition proportions: Using TYPE1: OLS

Scaled Condition Index	Scaled Proportion of the Variance of								
	Intercept	Age	Black	on any Diet	on Low- Calorie Diet	on Low-fat Diet	on Low-carb Diet	Calorie	Protein
1	<i>a</i>
2
3	<i>0.574</i>	.	.
3
3	<i>0.379</i>	.	.	.
4	.	.	<i>0.794</i>
5
6
8
9
11	.	.	.	<i>0.842</i>	<i>0.820</i>
12
22
26
38	<i>0.970</i>	<i>0.960</i>
157
581	<i>0.993</i>	<i>0.966</i>
Scaled Condition Index	Carbohydrate	Sugar	Dietary Fiber	Alcohol	Total Fat	Sat.fat ^b	Mono.fat ^c	Poly.fat ^d	
1
2
3
3
3
4
5
6
8
9
11
12
22
26	.	<i>0.633</i>
38
157	<i>0.304</i>	<i>0.866</i>	<i>0.890</i>	<i>0.904</i>	.
581	<i>0.988</i>	.	<i>0.482</i>	<i>0.986</i>	<i>0.696</i>

^a The scaled variance decomposition proportions smaller than 0.3 are omitted in this table.

^b Total Saturated Fatty Acids.

^c Total Monounsaturated Fatty Acids.

^d Total Polyunsaturated Fatty Acids.

Table 4
Scaled condition indexes and variance decomposition proportions: Using TYPE2: WLS

Scaled Condition Index	Intercept	Age	Black	Scaled Proportion of the Variance of				Calorie	Protein
				on any Diet	on Low-Calorie Diet	on Low-fat Diet	on Low-carb Diet		
1	^a
2
3	0.609	.	.
3
4	.	.	0.711	.	.	0.347	.	.	.
5
7
8
10
11	.	.	.	0.902	0.878
13
21
26
37	0.959	0.940
165
566	0.992	0.963

Scaled Condition Index	Carbohydrate	Sugar	Dietary Fiber	Alcohol	Total Fat	Sat.fat ^b	Mono.fat ^c	Poly.fat ^d
1
2
3
3
4
5
7
8
10
11
13
21
26	.	0.630
37
165	0.342	0.871	0.909	0.919
566	0.987	.	0.486	0.981	0.658	.	.	.

^a The scaled variance decomposition proportions smaller than 0.3 are omitted in this table.
^b Total Saturated Fatty Acids.
^c Total Monounsaturated Fatty Acids.
^d Total Polyunsaturated Fatty Acids.

Table 5
Scaled condition indexes and variance decomposition proportions: Using TYPE3: SWLS

Scaled Condition Index	Intercept	Age	Black	Scaled Proportion of the Variance of			Calorie	Protein
				on any Diet	on Low-Calorie Diet	on Low-fat Diet		
1	^a
2	.	.	.	0.717	1.278	0.553	.	.
3	0.697	.
3
3
4
5
7	0.766	1.686	0.461
8
10
11
13
21
26
37
165
566	0.318	1.095	1.190

^a The scaled variance decomposition proportions smaller than 0.3 are omitted in this table.
^b Total Saturated Fatty Acids.
^c Total Monounsaturated Fatty Acids.
^d Total Polyunsaturated Fatty Acids.

Table 5 (continued)
Scaled condition indexes and variance decomposition proportions: Using TYPE3: SWLS

Scaled Condition Index	Scaled Proportion of the Variance of							
	Carbohydrate	Sugar	Dietary Fiber	Alcohol	Total Fat	Sat.fat ^b	Mono.fat ^c	Poly.fat ^d
1
2
3
3
4
5
7
8
10
11
13
21
26	.	0.379
37	0.651	0.749	0.615
165	0.486	.	0.390
566	1.008	1.509	0.740	1.036	0.805			

^a The scaled variance decomposition proportions smaller than 0.3 are omitted in this table.
^b Total Saturated Fatty Acids.
^c Total Monounsaturated Fatty Acids.
^d Total Polyunsaturated Fatty Acids.

Table 6
Regression analysis output using TYPE3: SWLS

Variable	Original Model		Reduced Model	
	Coefficient	SE ^a	Coefficient	SE
Intercept	24.14*** ^b	2.77	24.20***	2.69
Age	0.06	0.08	0.06	0.08
Black	3.19***	1.04	3.67***	0.98
on any Diet ^c	1.79	1.52	1.28	1.80
on Low-calorie Diet	4.09**	1.50	4.59**	1.69
on Low-fat Diet	3.67	2.86	3.87	3.76
on Low-carb Diet	0.46	3.51	0.87	3.86
Calorie	-0.88	2.36		
Protein	7.05	9.59		
Carbohydrate	3.69	9.62		
Sugar	-0.31	1.11		
Dietary Fiber	-14.52*	5.89		
Alcohol	2.09	16.47		
Total Fat	29.34	31.37	1.47*	0.68
Total Saturated Fatty Acids	-15.90	20.18		
Total Monounsaturated Fatty Acids	-22.40	23.01		
Total Polyunsaturated Fatty Acids	-27.69	21.10		
Intracluster Coefficient ρ	0.0366		0.0396	

^a standard error.
^b p-value: *, 0.05; **, 0.01; ***, 0.005.
^c The reference category is “not being on diet” for all the on-diet variables here.

4.3 Study two: Reference level for categorical variables

As noted earlier, using non-survey data, dummy variables can also play an important role as a possible source for collinearity. The choice of reference level for a categorical variable may affect the degree of collinearity in the data. To be more specific, choosing a category that has a low frequency as the reference and omitting that level in order to

fit the model may give rise to collinearity with the intercept term. This phenomenon carries over to survey data analysis as we now illustrate.
We employed the four on-diet dummy variables used in the previous study, which we denote this section as “on any diet” (DIET), “on low-calorie diet” (CALDIET), “on low-fat diet” (FATDIET) and “one low-carbohydrate diet” (CARBDIET). The model considered here is:

$$\begin{aligned}
\text{BMI}_{hit} = & \beta_0 + \beta_{\text{black}} * \text{black}_{hit} \\
& + \beta_{\text{TOTAL.FAT}} * \text{TOTAL.FAT}_{hit} \\
& + \beta_{\text{DIET}} * \text{DIET}_{hit} \\
& + \beta_{\text{CALDIET}} * \text{CALDIET}_{hit} \\
& + \beta_{\text{FATDIET}} * \text{FATDIET}_{hit} \\
& + \beta_{\text{CARBDIET}} * \text{CARBDIET}_{hit} + \varepsilon_{hit} \quad (20)
\end{aligned}$$

where subscript *hit* stands for the t^{th} unit in the selected PSU *hi*, *black* is the dummy variable of black (*black* = 1 and non-black = 0), and *TOTAL.FAT* is the variable of daily total fat intake. According to the survey-weighted frequency table, 15.04% of the respondents are “on any diet”, 11.43% of them are “on low-calorie diet”, 1.33% of them are “on low-fat diet” and 0.47% of them are “on low-carbohydrate diet”. Being on a diet is, then, relatively rare in this example. If we choose the majority level, “not being on the diet”, as the reference category for all the four on-diet dummy variables, we expect no severe collinearity between dummy variables and the intercept, because most of values in the dummy variables will be zero. However, when fitting model (20), assume that an analyst is interested to see the impact of “not on any diet” on respondent’s BMI and reverses the reference level of variable *DIET* in model (20) into “being on the diet”. This change may cause a near dependency in the model because the column in **X** for variable *DIET* will nearly equal the column of ones for the

intercept. The following empirical study will illustrate the impact of this change on the regression coefficient estimation and how we should diagnose the severity of the resulting collinearity.

Table 7 and 8 present the regression analysis output of the model in (20) using the three regression types, OLS, WLS and SWLS, listed in Table 1. Table 7 is modeling the effects of on-diet factors on BMI by treating “not being on the diet” as the reference category for all the four on-diet variables. While Table 8 changes the reference level of variable *DIET* from “not on any diet” into “On any diet” and models the effect of “not on any diet” on BMI. The choice of reference level effects the sign of the estimated coefficient for variable *DIET* but not its absolute value or standard error. The size of the estimated intercept and its SE are different in Tables 7 and 8, but the estimable functions, like predictions, will of course, be the same with either set of reference levels. The SE of the intercept is about three times larger when “on any diet” is the reference level for variable *DIET* (Table 8) than when it is not (Table 7).

When choosing “not being on any diet” as the reference category for *DIET* in Table 9, the scaled condition indexes are relatively small and do not signify any remarkable near-dependency regardless of the type of regression. Only the last row for the largest condition index is printed in Tables 9 and 10. Often, the reference category for a categorical predictor will be chosen to be analytically meaningful. In this example, using “not being on any diet” would be logical.

Table 7
Regression analysis output: When “not on any diet” is the reference category for *DIET* variable in the model

Regression Type	Intercept	black	total.fat	on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1	27.22*** ^a	3.20***	0.95	3.03	1.75	2.75	-1.48
OLS	(0.61) ^b	(0.70)	(0.72)	(1.94)	(2.03)	(2.72)	(3.66)
TYPE2	26.13***	3.65***	1.44*	1.39	4.46*	3.86	0.94
WLS	(0.58)	(0.82)	(0.67)	(1.67)	(1.79)	(2.59)	(4.22)
TYPE3	26.13***	3.65***	1.44*	1.39	4.46**	3.86	0.94
SWLS	(0.64)	(0.99)	(0.63)	(1.80)	(1.70)	(3.73)	(3.87)

^a p-value: *, 0.05; **, 0.01; ***, 0.005.

^b Standard errors are in parentheses under parameter estimates.

Table 8
Regression analysis output: When “on any diet” is the reference category for *DIET* variable in the model

Regression Type	Intercept	black	total.fat	not on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1	30.25*** ^a	3.20***	0.95	-3.03	1.75	2.75	-1.48
OLS	(2.00) ^b	(0.70)	(0.72)	(1.94)	(2.03)	(2.72)	(3.66)
TYPE2	27.52***	3.65***	1.44*	-1.39	4.46*	3.86	0.94
WLS	(1.71)	(0.82)	(0.67)	(1.67)	(1.79)	(2.59)	(4.22)
TYPE3	27.52***	3.65***	1.44*	-1.39	4.46**	3.86	0.94
SWLS	(1.75)	(0.99)	(0.63)	(1.80)	(1.70)	(3.73)	(3.87)

^a p-value: *, 0.05; **, 0.01; ***, 0.005.

^b Standard errors are in parentheses under parameter estimates.

In Table 10, when “on any diet” is chosen as the reference category for variable DIET, the scaled condition indexes are increased and show a moderate degree of collinearity (condition index larger than 10) between the on-diet dummy variables and the intercept. Using the table of scaled variance decomposition proportions, in OLS and WLS, dummy variable for “not on any diet” and “on low-calorie diet” are involved in the dominant near-dependency with the intercept; however, in SWLS, only the dummy variable for “not on any diet” is involved in the dominant near-dependency with the intercept and the other three on-diet variables are much less worrisome.

5. Conclusion

Dependence between predictors in a linear regression model fitted with survey data affects the properties of parameter estimators. The problems are the same as for non-survey data: standard errors of slope estimators can be inflated and slope estimates can have illogical signs. In the extreme case when one column of the design matrix is exactly a linear combination of others, the estimating equations cannot be solved. The more interesting cases are ones where predictors are related but the dependence is not exact. The collinearity diagnostics that are available in standard software routines are not entirely appropriate for survey data. Any diagnostic that involves variance estimation needs

modification to account for sample features like stratification, clustering, and unequal weighting. This paper adapts condition numbers and variance decompositions, which can be used to identify cases of less than exact dependence, to be applicable for survey analysis.

A condition number of a survey-weighted design matrix $\mathbf{W}^{1/2}\mathbf{X}$ is the ratio of the maximum to the minimum eigenvalue of the matrix. The larger the condition number the more nearly singular is $\mathbf{X}^T\mathbf{W}\mathbf{X}$, the matrix which must be inverted when fitting a linear model. Large condition numbers are a symptom of some of the numerical problems associated with collinearity. The terms in the decomposition also involve “misspecification effects” if the model errors are not independent as would be the case in a sample with clustering. The variance of an estimator of a regression parameter can also be written as a sum of terms that involve the eigenvalues of $\mathbf{W}^{1/2}\mathbf{X}$. The variance decompositions for different parameter estimators can be used to identify predictors that are correlated with each other. After identifying which predictors are collinear, an analyst can decide whether the collinearity has serious enough effects on a fitted model that action should be taken. The simplest step is to drop one or more predictors, refit the model, and observe how estimates change. The tools we provide here allow this to be done in a way appropriate for survey-weighted regression models.

Table 9
Largest scaled condition indexes and its associated variance decomposition proportions: When “not on any diet” is the reference category for variable DIET in the model

Scaled Condition Index	Intercept	gender	total.fat	Scaled Proportion of the Variance of			
				on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1: OLS							
6	0.005	0.000	0.016	0.949	0.932	0.157	0.200
TYPE2: WLS							
6	0.013	0.008	0.020	0.938	0.926	0.189	0.175
TYPE3: SWLS							
6	0.006	0.007	0.013	0.686	0.741	0.027	0.061

Table 10
Largest scaled condition indexes and its associated variance decomposition proportions: When “on any diet” is the reference category for variable DIET in the model

Scaled Condition Index	Intercept	gender	total.fat	Scaled Proportion of the Variance of			
				not on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1: OLS							
17	0.982	0.001	0.034	0.968	0.831	0.155	0.186
TYPE2: WLS							
17	0.982	0.011	0.029	0.968	0.820	0.182	0.160
TYPE3: SWLS							
17	0.897	0.018	-0.006	0.971	0.318	0.014	-0.019

Acknowledgements

The authors thank the associate editor and referees whose comments led to important improvements. This work was partially supported by the U.S. National Science Foundation under Grant No. 0617081. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- Belsley, D.A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*, 38(2), 73-77.
- Belsley, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York: John Wiley & Sons, Inc.
- Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York: Wiley Interscience.
- Cook, R.D. (1984). Comment on demeaning conditioning diagnostics through centering. *The American Statistician*, 2, 78-79.
- Elliot, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 1, 23-34.
- Farrar, D.E., and Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.
- Fox, J. (1986). *Linear Statistical Models and Related Methods, with Applications to Social Research*. New York: John Wiley & Sons, Inc.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 1, 5-23.
- Hendrickx, J. (2010). *perturb: Tools for evaluating collinearity*. R package version 2.04. URL <http://CRAN.R-project.org/package=perturb>.
- Kish, L., and Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society B*, 36(1), 1-37.
- Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3341-3348.
- Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland.
- Li, J., and Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, 35, 1, 15-24.
- Li, J., and Valliant, R. (2011). Detecting groups of influential observations in linear regression using survey data-adapting the forward search method. Festschrift for Ken Brewer. *Pakistan Journal of Statistics*, 27, 507-528.
- Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Ph.D. thesis, University of Maryland.
- Liao, D., and Valliant, R. (2012). Variance inflation factors in the analysis of complex survey data. *Survey Methodology*, 38, 1, 53-62.
- Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.
- Marquardt, D.W. (1980). Comment on "A critique on some ridge regression methods" by G. Smith and F. Campbell: "You should standardize the predictor variables in your regression models". *Journal of the American Statistical Association*, 75(369), 87-91.
- Scott, A.J., and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77(380), 848-854.
- Silvey, S.D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society*, 31(3), 539-552.
- Snee, R.D., and Marquardt, D.W. (1984). Collinearity diagnostics depend on the domain of prediction, and model, and the data. *The American Statistician*, 2, 83-87.
- Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.
- Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons, Inc.
- Wissmann, M., Toutenburg, H. and Shalabh (2007). Role of categorical variables in multicollinearity in the linear regression model. Technical Report Number 008, Department of Statistics, University of Munich. Available at http://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf.
- Wood, F.S. (1984). Effect of centering on collinearity and interpretation of the constant. *The American Statistician*, 2, 88-90.

Bayesian inference for finite population quantiles from unequal probability samples

Qixuan Chen, Michael R. Elliott and Roderick J.A. Little¹

Abstract

This paper develops two Bayesian methods for inference about finite population quantiles of continuous survey variables from unequal probability sampling. The first method estimates cumulative distribution functions of the continuous survey variable by fitting a number of probit penalized spline regression models on the inclusion probabilities. The finite population quantiles are then obtained by inverting the estimated distribution function. This method is quite computationally demanding. The second method predicts non-sampled values by assuming a smoothly-varying relationship between the continuous survey variable and the probability of inclusion, by modeling both the mean function and the variance function using splines. The two Bayesian spline-model-based estimators yield a desirable balance between robustness and efficiency. Simulation studies show that both methods yield smaller root mean squared errors than the sample-weighted estimator and the ratio and difference estimators described by Rao, Kovar, and Mantel (RKM 1990), and are more robust to model misspecification than the regression through the origin model-based estimator described in Chambers and Dunstan (1986). When the sample size is small, the 95% credible intervals of the two new methods have closer to nominal confidence coverage than the sample-weighted estimator.

Key Words: Bayesian analysis; Cumulative distribution function; Heteroscedastic errors; Penalized spline regression; Survey samples.

1. Introduction

We consider inference for finite population quantiles of a continuous variable from a sample survey with unequal inclusion probabilities. The finite-population quantiles are usually estimated by the sample-weighted quantiles, a Horvitz-Thompson type estimator. Often in sample surveys the design variable (here, the inclusion probability) or a correlated auxiliary variable is measured on the non-sampled units, and this information can be used to improve the efficiency of the sample-weighted estimators (Zheng and Little 2003; Chen, Elliott, and Little 2010).

Methods for using auxiliary information in estimating finite-population distribution functions have been extensively studied. Chambers and Dunstan (1986) proposed a model-based method, illustrating their approach for a zero intercept linear regression superpopulation model. We refer to this estimator from now on as the CD estimator. Dorfman and Hall (1993) applied the CD approach, replacing the linear regression model with a non-parametric model. Lombardia, González-Manteiga, and Prada-Sánchez (2003, 2004) proposed a bootstrap approximation to these estimators based on resampling a smoothed version of the empirical distribution of the residuals. Kuk and Welsh (2001) also modified the CD approach to address departures from the model by estimating the conditional distribution of residuals as a function of the auxiliary variable. Rao, Kovar, and Mantel (RKM 1990) demonstrated advantages of

design-based ratio and difference estimators over the CD estimator when the model is misspecified. Wang and Dorfman (1996) suggested a weighted average of the CD and the RKM estimators. Kuk (1993) proposed a kernel-based estimator that combines the known distribution of the auxiliary variable with a kernel estimate of the conditional distribution of the survey variable given the value of the auxiliary variable. Chambers, Dorfman, and Wehrly (1993) proposed a kernel-smoothed model-based estimator, and Wu and Sitter (2001) and Harms and Duchesne (2006) proposed calibration type estimators.

Research on using auxiliary information for inference about finite population quantiles (defined as the inverse of the distribution function) is more limited. Chambers and Dunstan (1986) discussed estimation by inverting the CD estimator of the distribution function, but did not compare the performance of this quantile estimator with alternatives. Rao *et al.* (1990) proposed simple ratio and difference quantile estimators that were considerably more efficient than the sample-weighted estimator when the survey outcome was approximately proportional to the auxiliary variable.

We assume here unequal probability sampling with inclusion probabilities that are known for all the units in the population. We develop two Bayesian spline-model-based estimators of finite population quantiles that incorporate the inclusion probabilities. The first method is to estimate the distribution function at a number of sample values using

1. Qixuan Chen is Assistant Professor, Department of Biostatistics, Columbia University Mailman School of Public Health, 722 West 168 Street, New York, NY 10032. E-mail: qc2138@columbia.edu; Michael R. Elliott and Roderick J.A. Little are professors, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu and rlittle@umich.edu.

Bayesian penalized spline predictive estimators (Chen *et al.* 2010). The finite population quantiles are then estimated by inverting the predictive distribution function. The second method is a Bayesian two-moment penalized spline predictive estimator, which predicts the values of non-sampled units based on a normal model, with mean and variance both modeled with penalized splines on the inclusion probabilities. We compare the performance of these two new methods with the sample-weighted estimator, the CD estimator, and the RKM's ratio and difference estimators, using simulation studies on artificially generated data and farm survey data.

2. Estimators of the quantiles

Let s denote an unequal probability random sample of size n , drawn from the finite population of N identifiable units according to inclusion probabilities $\{\pi_i, i = 1, \dots, N\}$, which are assumed to be known for all the units before a sample is drawn. Let Y denote a continuous survey variable, with values $\{y_1, y_2, \dots, y_n\}$ observed in the random sample s . The finite-population α -quantile of Y is defined as:

$$\theta(\alpha) = \inf \left\{ t; N^{-1} \sum_{i=1}^N \Delta(t - y_i) \geq \alpha \right\}, \quad (1)$$

where $\Delta(u) = 1$ when $u \geq 0$ and $\Delta(u) = 0$ elsewhere. The $\theta(\alpha)$ is often estimated using the sample-weighted α -quantile $\hat{\theta}(\alpha) = \inf \{t; \hat{F}_w(t) \geq \alpha\}$, where $\hat{F}_w(t)$ is the sample-weighted distribution function given by

$$\hat{F}_w(t) = \sum_{i \in s} \pi_i^{-1} \Delta(t - y_i) / \sum_{i \in s} \pi_i^{-1}.$$

Woodruff (1952) proposed a method of calculating confidence limits for the sample weighted α -quantile. First, a pseudo-population is obtained by weighting each sample item by its sampling weight; the standard deviation of the percentage of items less than the estimated α -quantile is estimated; and the estimated standard deviation is multiplied by the appropriate z percentile and is added to and subtracted from α to construct the confidence limits for the percentage of items less than the estimated α -quantile. Finally, the values of the survey variable corresponding to the confidence limits of the percentage of items less than the estimated α -quantile are read-off the weighted pseudo-population arrayed in order of size. Variance estimation of the percentage of items in the pseudo-population less than the estimated α -quantile is discussed in Woodruff (1952). Sitter and Wu (2001) showed that the Woodruff intervals perform well even in moderate to extreme tail regions of the distribution function. An alternative variance estimate was derived by Francisco and Fuller (1991) using a smoothed version of the large-sample test inversion.

2.1 Bayesian model-based approach, inverting the estimated CDF

The finite population quantile function is the inverse of the finite population cumulative distribution function (CDF), defined as $F(t) = N^{-1} \sum_{i=1}^N \Delta(t - y_i)$, where $\Delta(x) = 1$ when $x \geq 0$ and $\Delta(x) = 0$ elsewhere. We can estimate the finite population quantiles by first building a continuous and strictly monotonic predictive estimate of $F(t)$, by treating $\Delta(t - y)$ as a binary outcome variable and applying methods for estimating finite population proportions.

In particular, Chen *et al.* (2010) proposed a Bayesian penalized spline predictive (BPSP) estimator for finite population proportions in unequal probability sampling. They regress the binary survey variable z on the inclusion probabilities in the sample, using the following probit penalized spline regression model (2) with m pre-selected fixed knots:

$$\Phi^{-1}(E(z_i | \beta, b, \pi_i)) = \beta_0 + \sum_{k=1}^p \beta_k \pi_i^k + \sum_{l=1}^m b_l (\pi_i - k_l)^p, \\ b_l \sim N(0, \tau^2). \quad (2)$$

Self-representing units are included by setting $\pi_i = 1$. Assuming non-informative prior distributions for β and τ^2 , they simulated draws of z for the non-sampled units from their posterior predictive distribution. A draw from the posterior distribution of the finite population proportion is then obtained by averaging the observed sample units and the draws of the non-sample units. This is repeated many times to simulate the posterior distribution of the finite population proportion. Simulation studies indicated that the BPSP estimator is more efficient than the sample-weighted and generalized regression estimators of the finite population proportion, with confidence coverage closer to nominal levels.

We employ the BPSP approach n times to estimate $F(t)$ at each of the sampled values of y , $t = \{y_1, y_2, \dots, y_n\}$. This estimator does not take into account the fact that we are estimating a whole distribution function, and is not necessarily a monotonic function. In addition, linear interpolation of the n estimated distribution functions may lead to a poorly-estimated CDF. To overcome these two problems, we fit a smooth cubic regression curve to the n estimated distribution functions with monotonicity constraints (Wood 1994). We denote the resulting estimated distribution function as $\hat{F}(t)$. The Bayesian model-based estimator of $\theta(\alpha)$, obtained by inverting the estimated CDF, is then defined as follows:

$$\hat{\theta}_{\text{inv-CDF}}(\alpha) = \inf \{t; \hat{F}(t) \geq \alpha\}. \quad (3)$$

We also fit two other monotonic smooth regression curves to the upper and lower limits of the 95% credible intervals (CI) of these estimated distribution functions, denoted as $\hat{F}_U(t)$ and $\hat{F}_L(t)$. To reduce computation time in our simulation studies, we only estimate the CDF at $k < n$ pre-selected sample points.

The basic idea behind this approach is shown graphically in Figure 1. Suppose a sample of size 100 is drawn from a finite population. We pick 20 observations from the sample and estimate their corresponding distribution functions and associated 95% CI using the BPSP estimator. In Figure 1(a) we plot the BPSP estimates of these 20 points with black dots and the upper and lower limits of 95% CI with “-” signs, and connect the upper and lower limits with solid lines. In Figure 1(b) we add three monotonic smooth predictive curves using black solid curve for the point estimate and black dash curves for the upper and lower limits of the 95% CI.

In Figure 1(c) we draw a horizontal line across the graph with α as the y-axis value. We read x_A , x , and x_B respectively from the x-axis such that $\hat{F}_L(x_A) = \alpha$, $\hat{F}(x) = \alpha$, and $\hat{F}_U(x_B) = \alpha$. Then x is the inverse-CDF Bayesian estimate of $\theta(\alpha)$. If the 95% CI of the distribution function $F(\cdot)$ is formed by splitting the tail areas of the posterior distribution equally, the interval formed by x_A and x_B is a 95% CI of $\theta(\alpha)$. The proof is as follows: If α is the lower limit of the 95% CI of $F(x_A)$, only 2.5 percent of the draws of $F(x_A)$ in the posterior distribution are smaller than α . That is,

$$\Pr(F^{-1}(\alpha) > F^{-1}(F(x_A))) \equiv \Pr(\theta(\alpha) > x_A) = 0.025.$$

Similarly with α as the upper limit of the 95% CI of $F(x_B)$, $\Pr(\theta(\alpha) < x_B) = 0.975$. Therefore, there is 95% probability that $\theta(\alpha)$ is within x_A and x_B in the posterior distribution, given the sample.

This inverse-CDF Bayesian model-based approach avoids strong modeling assumptions, and can be applied to normal or skewed distributions. Estimating the distribution function at all n sample units makes full use of the sample information, but is computationally intensive; estimating the distribution function at $k < n$ values reduces computation time at the expense of some loss of efficiency. In the traditional approach, the population quantiles are estimated by inverting the unsmoothed empirical CDF. We recommend fitting a smooth cubic regression curve to the estimated distribution functions before inverting the estimated CDF. The resulting quantile estimates are more efficient, because the smooth curve exploits information from all the data. Simulations not shown here suggest that the estimated CDF distribution function curve estimated based on a well-chosen subset of the k sample units is similar to the curve estimated based on all sample units, but the computation time is significantly reduced.

We suggest choosing the subset of k data points at evenly spaced intervals in the middle of the distribution, and more frequent intervals in the extremes to improve the estimate of the CDF in the tails. For instance, in our simulation study with a sample size of 100, we estimated the distribution functions at 20 points: the 3 smallest, the 3 largest, and 14 other equally spaced points in the middle of the ordered sample.

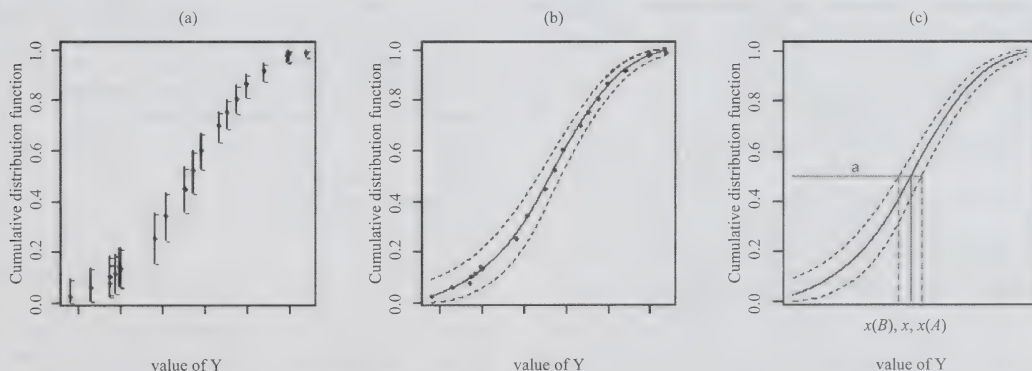


Figure 1 Inverse-CDF Bayesian model-based approach in estimating finite population distribution functions and associated quantiles illustrated using a sample of size 100 drawn from a finite population. (a) BPSP method is used to estimate the finite population distribution functions at 20 sample points; the dots denote BPSP estimators and the minus signs denote the upper and lower limits of the 95% CI. (b) Three monotonic smooth cubic regression models are fit on the BPSP estimators, upper limits, and lower limits; the solid curve is the predictive continuous distribution functions and the two dash curves are the 95% CI of the distribution functions. (c) The point estimate and 95% CI of population α -quantile are obtained by inverting the estimated CDF; x is the point estimate, and $x(B)$ and $x(A)$ are the lower and upper limits of the 95% CI

2.2 Bayesian two-moment penalized spline predictive approach

We consider alternative estimators of finite population quantiles of the form:

$$\tilde{\theta}(\alpha) = \inf \left\{ t; N^{-1} \left(\sum_{i \in s} \Delta(t - y_i) + \sum_{j \notin s} \Delta(t - \hat{y}_j) \right) \geq \alpha \right\}, \quad (4)$$

where \hat{y}_j is the predicted value of the j^{th} non-sample unit based on a regression on the inclusion probabilities $\{\pi_i\}$. A basic normal model for a continuous outcome assumes a mean function that is linear in $\{\pi_i\}$, that is:

$$Y_i \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 \pi_i, c_i \sigma^2), \quad (5)$$

with known constants c_i to model non-constant variance. This leads to a biased estimate of $\theta(\alpha)$ when the relationship is not linear. For estimating finite population totals, Zheng and Little (2003, 2005) replaced the linear mean function in (5) with a penalized spline, and assumed $c_i = \pi_i^{2k}$ with some known value of k . Simulations suggested that their model-based estimator of the finite population total outperforms the sample-weighted estimator, even when the variance structure is misspecified.

For estimation of quantiles rather than the total, correct specification of the variance structure is important in order to avoid bias. Therefore, we extend the penalized spline model in Zheng and Little (2003) by modeling both the mean and the variance using penalized splines. The two-moment penalized spline model can be written as (Ruppert, Wand, and Carroll 2003, page 264):

$$\begin{aligned} Y_i &\stackrel{\text{ind}}{\sim} N(\text{SPL}_1(\pi_i, k), \exp(\text{SPL}_2(\pi_i, k'))), \\ \text{SPL}_1(\pi_i, k) &= \beta_0 + \sum_{k=1}^p \beta_k \pi_i^k + \sum_{l=1}^{m_1} b_l (\pi_i - k_l)_+^p, \\ b_l &\stackrel{\text{iid}}{\sim} N(0, \tau_b^2), \\ \text{SPL}_2(\pi_i, k') &= \alpha_0 + \sum_{k=1}^p \alpha_k \pi_i^k + \sum_{l=1}^{m_2} v_l (\pi_i - k'_l)_+^p, \\ v_l &\stackrel{\text{iid}}{\sim} N(0, \tau_v^2). \end{aligned} \quad (6)$$

In (6), the mean and the logarithm of the variance are modeled as penalized splines (SPL_1) and (SPL_2) on $\{\pi_i\}$. Modeling the logarithm of the variance ensures positive estimates of the variance. We allow different numbers (m_1, m_2) and locations (k, k') of the knots for the two splines.

Ruppert *et al.* (2003) suggested an iterative approach to estimate the parameters in (6). They first assumed that SPL_2 was known and fitted a linear mixed model to estimate the parameters in SPL_1 . They calculated the square of the difference between Y and SPL_1 , which followed a Gamma distribution with the shape parameter as $1/2$ and the scale parameter of 2SPL_2 . They then fitted a generalized linear mixed model for the squared differences to estimate the parameters in SPL_2 . They iterated the above procedures until the parameter estimates converged. This iterative approach is simple to implement. However, our goal here is not to estimate the parameters but to obtain Bayesian predictions of Y for the non-sample units so that we can use (4) to estimate the quantiles.

Crainiceanu, Ruppert, Carroll, Joshi, and Goodner (2007) developed Bayesian inferential methodology for (6). They noted that the implementation of MCMC using multivariate Metropolis-Hastings steps is unstable with poor mixing properties. They suggested adding error terms to the second spline to make computations feasible, replacing sampling from complex full conditionals by simple univariate Metropolis-Hastings steps. This idea can be expressed as

$$Y_i \stackrel{\text{ind}}{\sim} N(\text{SPL}_1(\pi_i, k), \sigma_e^2(\pi_i)),$$

$$\log(\sigma_e^2(\pi_i)) \stackrel{\text{iid}}{\sim} N(\text{SPL}_2(\pi_i, k'), \sigma_A^2).$$

We used a prior distribution $N(0, 10^6)$ for the fixed effects parameters β and α , and a proper inverse-gamma prior distribution $\text{IGamma}(10^{-6}, 10^{-6})$ for the variance components τ_b^2 and τ_v^2 . We fixed the values of $\sigma_A^2 = 0.1$. The full conditionals of the posterior are detailed in Crainiceanu *et al.* (2007).

The posterior distribution of the finite population α -quantile is simulated by generating a large number D of draws and using the predictive estimator form

$$\tilde{\theta}^{(d)}(\alpha) = \inf \left\{ t; N^{-1} \left(\sum_{i \in s} \Delta(t - y_i) + \sum_{j \notin s} \Delta(t - \hat{y}_j^{(d)}) \right) \geq \alpha \right\},$$

where $\hat{y}_j^{(d)}$ is a draw from the posterior predictive distribution of the j^{th} non-sampled unit of the continuous outcome. The average of these draws simulates the Bayesian two-moment penalized spline predictive (B2PSP) estimator of the finite population α -quantile,

$$\hat{\theta}_{\text{B2PSP}}(\alpha) = D^{-1} \sum_{d=1}^D \tilde{\theta}^{(d)}(\alpha).$$

The Bayesian 95% credible interval for the population α -quantile in the simulations is formed by splitting the tail area equally between the upper and lower endpoints.

3. Simulation study

3.1 Simulation study with artificial data

We first simulated a super-population of size $M = 20,000$. The size variable X in the super-population takes 20,000 consecutive integer values from 710 to 20,709. A finite population of size $N = 2,000$ was then selected from the super-population using systematic probability proportional to size (pps) sampling with the probability proportional to the inverse of the size variable. Consequently, the size variable in the finite population has a right skewed distribution. The survey outcome Y was drawn from a normal distribution with mean $f(\pi)$ and error variance equal to 0.04 (homoscedastic error) or π (heteroscedastic error). Three different mean structures $f(\pi)$ were simulated: no association between Y and π (NULL) $f(\pi) = 0.5$, a linear association (LINUP) $f(\pi) = 6\pi$, and a nonlinear association (EXP) $f(\pi) = \exp(-4.64 + 52\pi)$. For each of the six simulation conditions, one thousand replicate finite populations were generated, and a systematic pps sample ($n = 100$) was drawn from each population with x as the size variable; thus $\pi_i = nx_i / \sum_{j=1}^N x_j$. Scatter plots of Y versus π for these six populations are displayed in Figure 2.

We compared the performance of the Bayesian inverse-CDF and the B2PSP estimators with five alternative approaches:

- SW, the sample-weighted estimator defined by inverting \hat{F}_w .
- Smooth-SW, the smooth sample-weighted estimator. A smooth cubic regression curve was fit to \hat{F}_w , and denoted as \tilde{F}_w . The smooth sample-weighted estimator is then defined as $\tilde{\theta}_w = \inf\{t; \tilde{F}_w \geq \alpha\}$.
- CD, the Chambers and Dunstan estimator (1986), by assuming the following model: $Y_i = \beta\pi_i + \pi_i U_i$, where U_i is an independent and identically distributed random variable with zero mean.
- Ratio, the RKM's ratio estimator (1990) given by $\{\hat{\theta}_y(\alpha) / \hat{\theta}_x(\alpha)\} \times \theta_x(\alpha)$, where $\hat{\theta}_y(\alpha)$ and $\hat{\theta}_x(\alpha)$ denotes respectively the sample-weighted estimates for Y and the size variable X , and $\theta_x(\alpha)$ is the known population quantile of X .
- Diff, the RKM's difference estimator (1990) given by $\hat{\theta}_y(\alpha) + \hat{R} \times \{\theta_x(\alpha) - \hat{\theta}_x(\alpha)\}$, where \hat{R} is the sample-weighted estimate of Y/X .

The seven estimators for the finite-population 10th, 25th, 50th, 75th, and 90th percentiles were compared in terms of

empirical bias and root mean squared error (RMSE). Because of the complexity in the variance estimation for the CD and RKM's estimators, we only compared the average width and the non-coverage rate of the 95% confidence/credible interval (CI) for the two Bayesian model-based estimators and the sample-weighted estimator. For the 95% CI, we used Woodruff's method for the sample-weighted estimator, the method illustrated in Figure 1(c) for the inverse-CDF Bayesian estimator, and the 95% posterior probability of the quantile with equal tails for the B2PSP estimator. We used cubic splines with 15 equally spaced knots.

Tables 1 and 2 show the empirical bias and RMSE for the three normal distributions with homoscedastic errors and with heteroscedastic errors, respectively. Overall, the empirical bias in estimating the five quantiles is similar using the two Bayesian estimators, the two sample-weighted estimators, and the RKM's two design-based estimators. In contrast, the CD estimator has large bias and RMSE in all scenarios except for LINUP with heteroscedastic error, where its underlying model is correctly specified. The two Bayesian model-based estimators yield smaller root mean squared errors than the other estimators, and this improvement in efficiency is substantial in some scenarios, especially using the B2PSP estimator. By applying a smooth cubic regression curve on the estimated empirical sample-weighted CDF, the smooth-sample-weighted estimator gains some efficiency over the conventional sample-weighted estimators, but the RMSE is still larger than the Bayesian Inverse-CDF estimator. Comparisons of the three design-based estimators suggest that none of the three estimators uniformly dominates the other two. Specifically, the sample-weighted estimator has smaller RMSE than the RKM difference and ratio estimators for all five quantiles in the NULL and for the lower quantiles in the LINUP and EXP populations; on the other hand, the RKM estimators have smaller RMSE at the upper quantiles in the LINUP and EXP populations.

Table 3 shows the average width and non-coverage rate of 95% CI for the two Bayesian model-based estimators and the sample-weighted estimator. Overall, the two Bayesian model-based estimators yield shorter average 95% CI widths than the sample-weighted estimator. The coverage rate of the 95% CI is similar among the three estimators, except that when α is equal to 0.1, where the 95% CI of the B2PSP estimator has the shortest average width and very good coverage, while the sample-weighted estimator has serious under-coverage. This happens because the Woodruff method for estimating the variance of the sample-weighted estimator is based on a large sample assumption, but here the pps sampling leads to only a small number of cases being sampled in the lower tail.

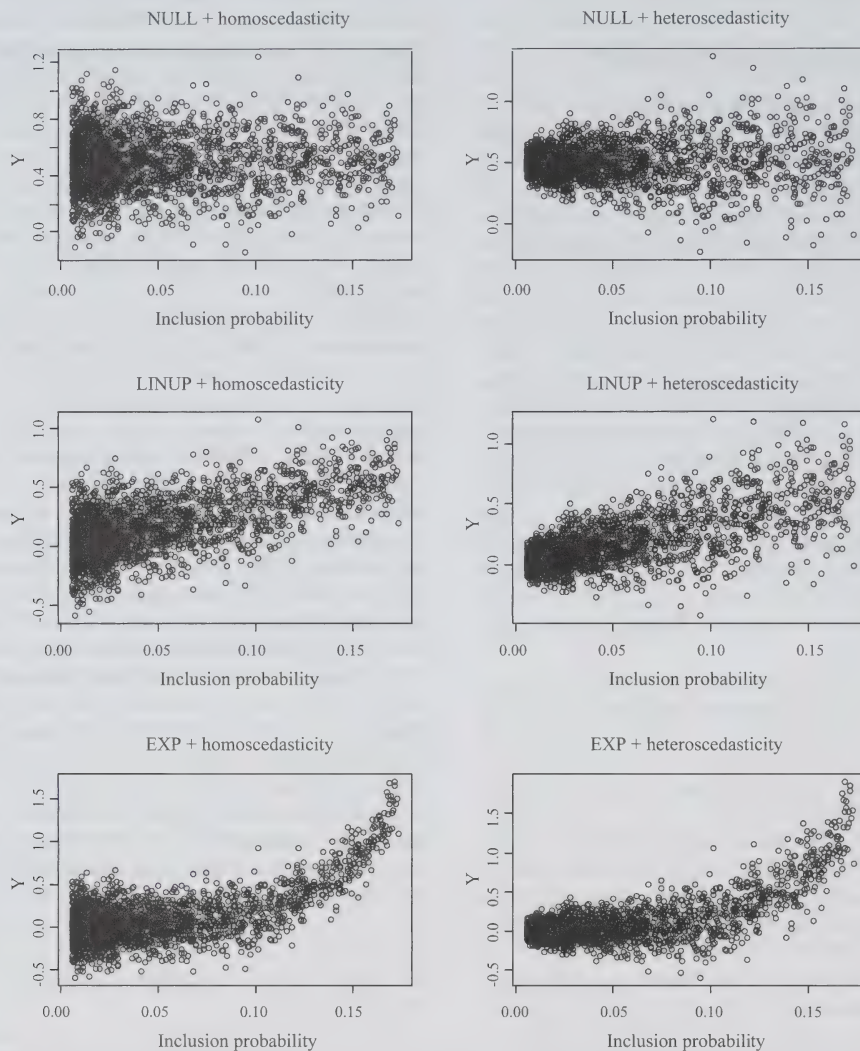


Figure 2 Scatter plots of Y versus the inclusion probabilities for the six artificial finite populations of size equal to 2,000

Although the sample-weighted estimator performs similarly with the two Bayesian spline-model-based estimators in terms of overall empirical bias, the conditional bias of estimates varies largely as the sample mean of the inclusion probability increases. Following Royall and Cumberland (1981), the estimates from the 1,000 samples were ordered according to the sample mean of the inclusion probabilities and were split into 20 groups of 50 each, and then the empirical bias was calculated for each group. Figure 3

displays the conditional bias of the two Bayesian estimators and the sample-weighted estimator for the 90th percentile in the “EXP + homoscedastic error” case. Figure 3 shows that there is a linear trend for the bias in the sample-weighted estimator as the sample mean of the inclusion probabilities increases, while the grouped bias of the two Bayesian spline-model-based estimators is less affected by the sample mean of inclusion probabilities. Similar findings are also seen in other scenarios.

Table 1
Comparisons of empirical bias and root mean squared errors $\times 10^3$ of $\theta(\alpha)$ for $\alpha = 0.1, 0.25, 0.5, 0.75$, and 0.9 : Scenarios with homoscedastic errors

	Empirical bias					Empirical RMSE				
	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
<i>NULL</i>										
Inverse-CDF	-6	-3	-1	-1	-5	46	37	36	37	45
B2PSP	-5	-1	1	2	6	41	33	31	34	42
SW	-5	-3	-1	-4	-6	54	41	39	41	50
Smooth-SW	-7	-4	-1	-2	-5	50	39	37	38	47
CD	-197	-272	-265	-108	168	203	274	266	115	189
RKM's Ratio	3	25	33	16	6	77	125	159	112	79
RKM's Diff	-5	-1	6	14	14	58	58	94	122	113
<i>LINUP</i>										
Inverse-CDF	-15	-3	-2	-1	-2	70	49	39	34	33
B2PSP	-3	-1	1	4	7	56	43	35	31	29
SW	-15	-3	-3	-2	-6	77	57	48	44	42
Smooth-SW	-14	-5	-2	-1	-4	72	53	45	42	41
CD	101	35	-37	-49	1	104	38	39	53	31
RKM's Ratio	-23	-9	2	5	-0.2	95	67	53	51	40
RKM's Diff	-15	-4	-4	-0.2	-2	77	55	45	43	38
<i>EXP</i>										
Inverse-CDF	-8	0.4	4	7	4	60	45	41	43	49
B2PSP	-10	-6	-3	0.3	13	52	40	35	36	36
SW	-9	-3	-2	-2	-8	65	49	46	50	72
Smooth-SW	-12	-5	-2	-1	-2	62	47	43	46	68
CD	92	54	14	19	61	96	57	21	31	75
RKM's Ratio	-17	-11	1	3	-5	87	65	50	53	55
RKM's Diff	-9	-4	-2	-2	-7	65	49	47	47	59

Table 2
Comparisons of empirical bias and root mean squared errors $\times 10^3$ of $\theta(\alpha)$ for $\alpha = 0.1, 0.25, 0.5, 0.75$, and 0.9 : Scenarios with heteroscedastic errors

	Empirical bias					Empirical RMSE				
	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
<i>NULL</i>										
Inverse-CDF	-9	-8	-2	4	1	30	24	22	24	31
B2PSP	-6	-6	1	7	7	25	21	19	23	27
SW	-4	-3	-2	-1	-5	34	26	23	26	35
Smooth-SW	-4	-5	-2	1	-4	34	26	23	26	35
CD	-298	-325	-253	-46	270	302	327	255	60	288
RKM's Ratio	8	31	32	16	5	81	143	154	94	57
RKM's Diff	-5	-1	6	17	16	44	54	87	113	97
<i>LINUP</i>										
Inverse-CDF	-11	-1	5	2	-3	32	24	24	29	35
B2PSP	-10	-1	7	3	1	29	22	22	24	30
SW	-5	-1	-0.1	-1	-4	31	28	33	45	51
Smooth-SW	-11	-3	2	-0.4	-5	32	26	30	44	50
CD	10	7	6	7	11	20	13	13	20	32
RKM's Ratio	-7	-3	2	3	1	36	29	30	35	41
RKM's Diff	-5	-2	-1	1	-0.2	32	27	28	33	41
<i>EXP</i>										
Inverse-CDF	-8	-3	5	7	-3	30	23	23	30	48
B2PSP	-11	-7	2	6	7	28	23	20	25	36
SW	-3	-3	-2	1	-2	30	26	26	41	84
Smooth-SW	-8	-5	1	2	-5	30	23	24	39	86
CD	18	16	35	84	68	27	21	38	88	81
RKM's Ratio	-5	-6	-1	2	-0.1	36	31	27	32	62
RKM's Diff	-3	-3	-2	1	-0.1	32	28	28	31	67

Table 3
Comparisons of average width and non-coverage rate of 95% CI $\times 10^3$ of $\theta(\alpha)$ for $\alpha = 0.1, 0.25, 0.5, 0.75$, and 0.9

	Average width of 95% CI					Non-coverage rate of 95% CI				
	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
<i>Homoscedastic errors</i>										
<i>NULL</i>										
Inverse-CDF	199	156	141	152	184	46	35	44	38	67
B2PSP	178	134	118	134	177	52	55	61	59	50
SW	195	164	151	167	237	112	65	46	40	38
<i>LINUP</i>										
Inverse-CDF	257	207	157	139	141	61	45	37	46	52
B2PSP	230	167	134	123	121	58	54	44	57	59
SW	248	231	188	179	187	119	60	42	41	39
<i>EXP</i>										
Inverse-CDF	234	184	163	177	234	59	44	47	40	42
B2PSP	217	157	132	144	156	54	59	55	53	60
SW	231	199	175	210	402	106	64	47	40	40
<i>Heteroscedastic errors</i>										
<i>NULL</i>										
Inverse-CDF	146	104	90	101	137	42	43	38	38	47
B2PSP	107	89	79	89	107	38	49	37	68	65
SW	146	101	91	113	169	80	60	51	37	42
<i>LINUP</i>										
Inverse-CDF	131	107	104	124	154	70	31	36	42	40
B2PSP	125	97	87	93	116	47	35	50	58	52
SW	141	110	133	184	219	138	69	41	50	42
<i>EXP</i>										
Inverse-CDF	131	99	99	134	242	63	49	34	40	41
B2PSP	116	92	84	98	139	57	55	40	63	59
SW	135	100	106	186	378	111	65	46	45	34

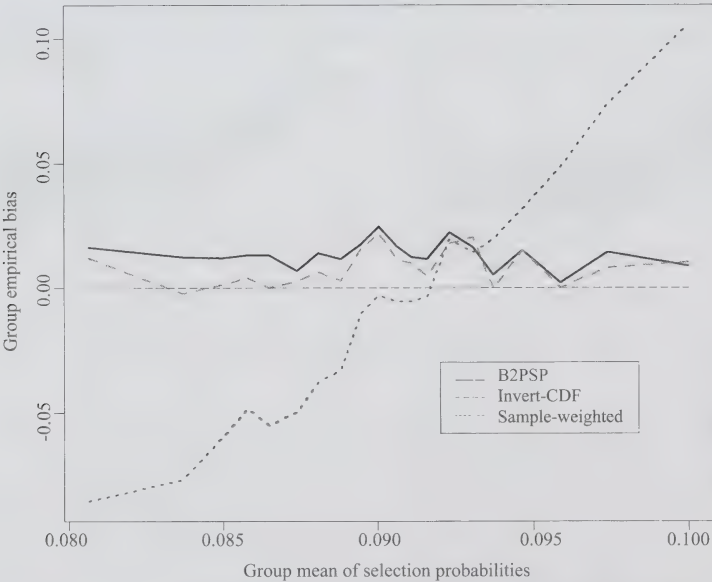


Figure 3 Variation of empirical bias of the three estimators for 90th percentile from the “EXP + homoscedasticity” case

3.2 Simulation study with the broadacre farm survey data

The B2PSP estimator assumes the outcome has a normal distribution, after conditioning on the inclusion probabilities. Since the inverse-CDF Bayesian model-based approach does not assume normality, we might expect it to out-perform the B2PSP when the normality assumption is violated. This motivates a comparison of the sample-weighted and the inverse-CDF Bayesian estimators for non-normal data.

The population considered here is defined by 398 broadacre farms (farms involved in the production of cereal crops, beef, sheep and wool) with 6,000 or less hectares that participated in the 1982 Australian Agricultural and Grazing Industries Survey carried out by the Australian Bureau of Agricultural and Resource Economics (ABARE 2003). The Y variable is the total farm cash receipts. One thousand systematic pps samples of size equal to 100 were drawn with the farm area, X , as the size variable, that is, larger farms are more likely to be selected into the sample. Figure 4 is the scatter plot of Y versus the size variable X for these

farms, with filled circles representing a selected pps sample. This shows that the variation of Y increases as X increases. Moreover, Y is right-skewed given X . A simulation study using this broadacre farms data was conducted to compare the two Bayesian spline-model-based estimators with the sample-weighted estimator.

Table 4 shows the simulation results. The inverse-CDF Bayesian approach yields smaller empirical bias and RMSE, and shorter average length of 95% CI than the sample-weighted estimator in general. The 95% CI of the inverse-CDF Bayesian approach also have closer to nominal level confidence coverage than the sample-weighted estimator when α is 0.1 and 0.25. However, in the upper tail with $\alpha = 0.90$, the non-coverage rate of the inverse-CDF Bayesian approach is higher than the nominal level 0.05, while the Woodruff CI of the sample-weighted estimator does well. This is consistent with the findings of Sitter and Wu (2001) that the Woodruff intervals perform well even in the moderate to extreme tail regions of the distribution function. Since the conditional normality assumption is not reasonable here, the B2PSP estimator is biased and the 95% CI has poor confidence coverage.

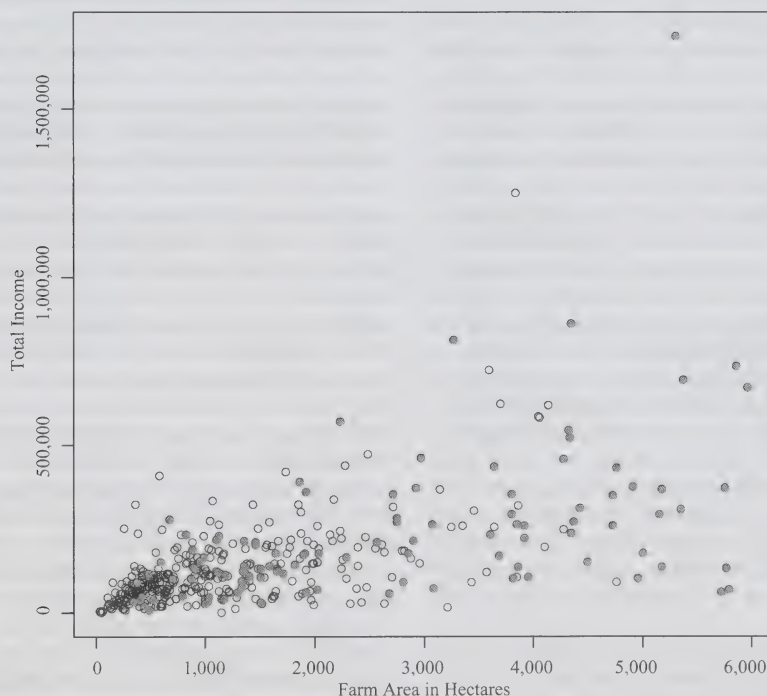


Figure 4 Scatter plot of the broadacre farm data with the filled circles representing a pps sample

Table 4
Empirical bias $\times 10^{-2}$, root mean squared errors $\times 10^{-2}$, average width of 95% CI $\times 10^{-2}$, and non-coverage rate of 95% CI $\times 10^3$ of $\theta(\alpha)$ for $\alpha = 0.1, 0.25, 0.5, 0.75$, and 0.9 : The broadacre farm data

	0.1	0.25	0.5	0.75	0.9
<i>Empirical bias</i>					
Inverse-CDF	8	14	10	-22	-60
B2PSP	-110	-125	-63	-12	88
SW	20	-19	-17	-21	-61
<i>Empirical RMSE</i>					
Inverse-CDF	117	117	108	164	256
B2PSP	113	141	124	140	206
SW	132	173	167	226	350
<i>Average width of 95% CI</i>					
Inverse-CDF	402	443	501	697	906
B2PSP	170	327	539	726	964
SW	285	468	615	864	1,589
<i>Non-coverage rate of 95% CI</i>					
Inverse-CDF	96	53	26	52	90
B2PSP	670	258	42	8	17
SW	220	121	68	42	44

4. Discussion

Sample-weighted estimators for finite population quantiles are widely used in survey practice. Although the sample-weighted estimators with Woodruff’s confidence intervals are easy to compute and can provide valid large-sample inferences, they may be inefficient and confidence coverage can be poor in small-to-moderate-sized samples. Model-based estimators can improve the efficiency of the estimates when the model is correctly specified, but lead to biased estimates when the model is misspecified. To achieve the balance between robustness and efficiency, we considered spline-model-based estimators. For the quantile estimation of a continuous survey variable, we can either estimate the model-based distribution functions and invert the distribution functions to obtain quantiles, or model the survey outcome on the inclusion probabilities directly. In this paper, we proposed two Bayesian spline-model-based quantile estimators. The first method is the Bayesian inverse-CDF estimator, obtained by inverting the spline-model-based estimates of distribution functions. The second method is the B2PSP estimator, estimated by assuming a normal distribution for the continuous survey outcome, with the mean function and the variance function both modeled using splines.

The simulations suggest that the two Bayesian spline-model-based estimators outperform the sample-weighted estimator, the design-based ratio and difference estimators, as well as the CD model-based estimator when its assumed model is incorrect. Both new methods yield smaller root

mean squared errors whether there is no association, a linear association, or a nonlinear association between the survey outcome and the inclusion probability. In some scenarios, the improvement in efficiency using the two Bayesian methods is substantial. When the normality assumption of the survey outcome given the inclusion probabilities is true, the B2PSP estimator has smaller RMSE and shorter credible interval than the inverse-CDF approach. Moreover, the two Bayesian model-based estimators are robust to the misspecification in both the mean and variance functions. In contrast, the CD model-based estimator is biased and inefficient when either the mean function or the variance function is misspecified. Finally, the Bayesian model-based methods have the advantage of easier calculation of the 95% CI and inference based on the posterior distributions of parameters. This is appealing, because variance estimation for the alternative design-based estimators can be complicated. Woodruff’s variance estimation method for sample-weighted estimator performs well when a large fraction of the data is selected from the finite population, even in the moderate to extreme tail regions of the distribution function. However, when data from the population is sparse, the Woodruff’s method tends to underestimate the confidence coverage, whereas both Bayesian methods have closer to nominal level confidence coverages.

All the three design-based estimators have comparable overall empirical bias to the two Bayesian spline-model-based estimators. However, there is a linear trend in the variation of bias for the sample-weighted estimator as the sample mean of inclusion probabilities increases. When

there is no association between the survey outcome and the inclusion probability, the ratio and difference estimators have relatively larger bias and RMSE than the sample-weighted estimator. However, in some simulation scenarios, the ratio and difference estimators achieve smaller RMSE than the sample-weighted estimator. The comparison between the conventional sample-weighted estimator and the smooth sample-weighted estimator suggests that fitting a smooth cubic curve to the sample-weighted CDF can improve the efficiency, but the smooth sample-weighted estimator still has larger RMSE than the Bayesian inverse-CDF estimator.

For normally distributed data, we recommend the use of the B2PSP estimator over the other estimators, because of smaller bias, smaller RMSE, and better confidence coverage with shorter interval length. The B2PSP estimator and its 95% posterior probability interval are easy to obtain using the algorithm proposed by Crainiceanu *et al.* (2007), which also has the advantage of relatively short computation time.

The B2PSP estimator is potentially biased when the conditional normal assumption does not hold. One possibility here is to transform the survey outcome to make the conditional normality assumption more reasonable. The B2PSP estimator can be applied to the transformed data, and the draws from the posterior distributions of the non-sampled units are transformed back to the original scale before estimating the quantiles of interest.

In our simulations with non-normal data, the inverse-CDF Bayesian approach was still more efficient than the sample-weighted estimator. Improvement in the confidence coverage was restricted to situations where the sample size is small, with Woodruff's CI method performing well when the large sample assumption holds. Thus for non-normal data where there no clear transformation to improve normality, we do not recommend the inverse-CDF Bayesian approach when the sample size is large. Given the good properties of the B2PSP estimator in the normal setting, one extension for future work is to relax the normality assumption in our proposed approaches.

We use the probability of inclusion as the auxiliary variable here. When there is only one relevant auxiliary variable, it does not matter whether the inclusion probability or the auxiliary variable is modeled. However, if there is more than one relevant auxiliary variable, the inclusion probability is the key auxiliary variable that needs to be modeled corrected, since misspecification of the model relating the survey outcome to the inclusion probability leads to bias. When other auxiliary variables are observed for all the units in the finite population, both of our Bayesian estimators can be easily extended to include additional auxiliary covariates by adding linear terms for these variables in the corresponding penalized spline model.

One reviewer suggested an alternative weighted Dirichlet approach, which is simple to calculate but it does not utilize the known auxiliary variables in the non-sampled units. Another possibility is to re-define the CD estimator by using the spline model we have used to define the B2PSP. Specifically, instead of assuming a regression model through the origin, a spline model is fitted to the first and second order moments of the conditional distribution of survey outcome given the inclusion probability. The spline-based CD estimator should perform similarly to the B2PSP estimator, and its variance can be estimated using resampling methods.

In the official statistics context, the methods in this article illustrate the potential benefits of a paradigm shift from design-based methods towards Bayesian modeling that is geared to yielding inferences with good frequentist properties. Design-based statistical colleagues raise two principal objections to this viewpoint.

First, the idea of an overtly model-based - even worse, Bayesian - approach to probability surveys is not well received, although our emphasis here is on Bayesian methods with good randomization properties. We believe that classical design-based methods do not provide the comprehensive approach needed for the complex problems that increasingly arise in official statistics. Judicious choices of well-calibrated models are needed to tackle such problems. Attention to design features and objective priors can yield Bayesian inferences that avoid subjectivity, and modeling assumptions are explicit, and hence capable of criticism and refinement. See Little (2004, 2012) for more discussion of these points.

The second objection is that Bayesian methods are too complex computationally for the official statistics world, where large number of routine statistics need to be computed correctly and created in a timely fashion. It is true that current Bayesian computation may seem forbidding to statisticians familiar with simple weighted statistics and replicate variance methods. Sedransk (2008), in an article strongly supportive of Bayesian approaches, points to the practical computational challenges as an inhibiting feature. We agree that work remains to meet this objection, but we do not view it insuperable. Research on Bayesian computation methods has exploded in recent decades, as have our computational capabilities. Bayesian models have been fitted to very large and complex problems, in some cases much more complex than those typically faced in the official statistics world.

Acknowledgements

We thank Dr. Philip Kokic in the Commonwealth Scientific and Industrial Research Organisation for providing us the broadacre form data. We also thank an associate editor

and referees for their helpful comments on the original version of this paper.

References

- ABARE (2003). Australian farm surveys report 2003. Canberra.
- Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of American Statistical Association*, 88, 268-277.
- Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chen, Q., Elliott, M.R. and Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36, 1, 23-34.
- Crainiceanu, C.M., Ruppert, D., Carroll, R.J., Joshi, A. and Goodner, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic error. *Journal of Computational and Graphical Statistics*, 16, 265-288.
- Dorfman, H., and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1474.
- Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.
- Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 1, 37-52.
- Kuk, A.Y.C. (1993). A kernel method for estimating finite population functions using auxiliary information. *Biometrika*, 80, 385-392.
- Kuk, A.Y.C., and Welsh, A.H. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society, Series B*, 63, 277-292.
- Little, R.J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, DOI: 10.1198/016214504000000467. {70}, 99, 546-556.
- Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion and rejoinder). *Journal of Official Statistics*, 28, 309-334.
- Lombardía, M.J., González-Manteiga, W. and Prada-Sánchez, J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference*, 116, 367-388.
- Lombardía, M.J., González-Manteiga, W. and Prada-Sánchez, J.M. (2004). Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimate of a finite population distribution function. *Journal of Nonparametric Statistics*, 16, 63-90.
- Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution function and quantile from survey data using auxiliary information. *Biometrika*, 77, 365-375.
- Royall, R.M., and Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - An empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.
- Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24, 495-506.
- Sitter, R.R., and Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters*, 52, 353-358.
- Wang, S., and Dorfman, A.H. (1996). A new estimator for the finite population distribution function. *Biometrika*, 83, 639-652.
- Wood, S.N. (1994). Monotonic smoothing splines fitted by cross validation SIAM. *Journal on Scientific Computing*, 15, 1126-1133.
- Woodruff, R. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.
- Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complex auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.
- Zheng, H., and Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.
- Zheng, H., and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Multiple imputation with census data

Satkartar K. Kinney¹

Abstract

A benefit of multiple imputation is that it allows users to make valid inferences using standard methods with simple combining rules. Existing combining rules for multivariate hypothesis tests fail when the sampling error is zero. This paper proposes modified tests for use with finite population analyses of multiply imputed census data for the applications of disclosure limitation and missing data and evaluates their frequentist properties through simulation.

Key Words: Finite Populations; Missing data; Significance testing; Synthetic data.

1. Introduction

Multiple imputation was first proposed for handling non-response in large complex surveys (Rubin 1987). Several other uses for multiple imputation have since been proposed, including statistical disclosure limitation and measurement error. An appeal of multiple imputation is that standard methods can be applied to each imputed dataset and then simple combining rules applied, which vary between applications. See Reiter and Raghunathan (2007) for a detailed overview of the different rules and applications. Existing multiple imputation combining rules were developed for use with random samples and superpopulation models (Deming and Stephan 1941). In finite population analyses of census data, where the sampling variance is zero, the combining rules for univariate estimands can still be applied as a special case; however, hypothesis tests for multivariate estimands break down.

Motivated by the use of multiple imputation to generate partially synthetic data (Rubin 1993; Little 1993) for the U.S. Census Bureau's Longitudinal Business Database (Kinney, Reiter, Reznick, Miranda, Jarmin and Abowd 2011), an economic census, this paper derives a multivariate test for finite populations for use with partially synthetic data and extends it to the application of missing data. Extensions to other multiple imputation applications are expected to be straightforward.

The remainder of this paper is organized as follows. Section 2 describes the case of partially synthetic data and Section 3 presents the extension to missing data. Simulations in Section 4 evaluate the combining rules for both the missing data and partially synthetic data cases.

2. Partially synthetic data

Partially synthetic datasets are constructed by replacing selected values in the confidential data with m independent draws from their posterior predictive distribution. For a

finite population of size N , let $Z_j = 1, j = 1, \dots, N$ indicate that unit j has been selected to have any observed values replaced with imputations. Imputations should only be made from the posterior predictive distribution of those units with $Z_j = 1$. For simplicity, in this paper, we assume $Z_j = 1, j = 1, \dots, N$. Let $Y = (y_1, \dots, y_d)$ be the matrix of confidential variables that will be replaced with imputations and X the matrix of variables that will not be replaced. Let $D_{\text{cen}} = (X, Y)$ represent a census of all N units containing confidential data and assume that all units are fully observed, i.e., no missing values are present. Let $Y_{\text{rep}}^{(i)}, i = 1, \dots, m$ be the i^{th} imputation of Y , and let $D_{\text{syn}}^{(i)} = (X, Y_{\text{rep}}^{(i)})$. The set $D_{\text{syn}} = \{D_{\text{syn}}^{(i)}, i = 1, \dots, m\}$ is what is released to the public.

Any proper imputation procedure from the broad literature on multiple imputation may be used to generate D_{syn} from D_{cen} . The finite population methods proposed here can be used regardless of whether a finite population was assumed in the generation of D_{syn} . Under a finite population assumption, since the data are a fully observed census the imputation model parameters would be considered known and fixed. See Reiter and Kinney (2012) for an illustration of how valid inferences are obtained from partially synthetic random samples generated with both fixed and random imputation model parameters. Simulations (not shown) confirm the same is true in the finite population case.

An analyst with access to D_{syn} but not D_{cen} can obtain valid inferences for a scalar or vector estimand Q using the following quantities:

$$\bar{Q}_m = \frac{1}{m} \sum_{i=1}^m Q^{(i)} \quad (2.1)$$

$$\bar{U}_m = \frac{1}{m} \sum_{i=1}^m U^{(i)} \quad (2.2)$$

$$B_m = \frac{1}{m-1} \sum_{i=1}^m (Q^{(i)} - \bar{Q}_m) (Q^{(i)} - \bar{Q}_m)' \quad (2.3)$$

where $Q^{(i)}$, $i = 1, \dots, m$, is the point estimate of Q obtained from $D_{\text{syn}}^{(i)}$, $U^{(i)}$ is the estimated variance of Q , and B_m is the sample variance of the $Q^{(i)}$, $i = 1, \dots, m$.

When there is no sampling variance the combining rules for scalar Q derived by Reiter (2003) can be applied as a special case where $\bar{U}_m = 0$. The resulting simplification means the approximations of Reiter (2003) are not needed and the exact posterior under multivariate normal theory is $(Q | D_{\text{syn}}) \sim t_{m-1}(\bar{Q}_m, B_m/m)$. For a vector Q , however, the hypothesis test of Reiter (2005) relies on the assumption that B_∞ is proportional to \bar{U}_∞ , i.e., the proportion of information replaced with imputations is the same across components of Q , so a different assumption is needed for the case $\bar{U}_\infty = 0$.

2.1 Proposed multivariate test

In this section an alternate test is derived based on the stronger assumption that $B_\infty = r_\infty I$, for a scalar quantity r_∞ and k -dimensional identity matrix I . In other words, the between-imputation variance is constant across components of Q , and B_∞ is assumed to be diagonal. In both the Reiter (2005) test and the proposed test, one averages across variance components so the test is moderately robust to this assumption; however, the randomization validity declines when the estimates of Q , $Q^{(i)}$, $i = 1, \dots, m$, are highly correlated. This is evaluated with simulations in Section 4.3. Comparable tests based on the assumption $B_\infty \propto \bar{U}_\infty$ are known to lose power when the assumption is not met (Li *et al.* 1991).

The proposed test for the hypothesis $H_0: Q = Q_0$ is conducted by referring the test statistic

$$S_c = \frac{(Q_0 - \bar{Q}_m)'(Q_0 - \bar{Q}_m)}{kr_c}$$

to an $F_{k, k(m-1)}$ distribution, where $r_c = 1/m \text{tr}(B_m)/k$.

Under the assumption $B_\infty = r_\infty I$, the Bayesian p -value is given by

$$\begin{aligned} & \int P(\chi_k^2 > (Q_0 - \bar{Q})' T_\infty^{-1} (Q_0 - \bar{Q}) | D_{\text{syn}}, B_\infty) \\ & \quad P(B_\infty | D_{\text{syn}}) dB_\infty \quad (2.4) \\ & = \int P\left(\chi_k^2 > \frac{(Q_0 - \bar{Q})' I (Q_0 - \bar{Q})}{r_\infty / m} \mid D_{\text{syn}}, r_\infty\right) \\ & \quad P(r_\infty | D_{\text{syn}}) dr_\infty \\ & = \int P\left(\frac{\chi_k^2}{k} \cdot \frac{r_\infty}{mr_c} > S_c \mid D_{\text{syn}}, r_\infty\right) \\ & \quad P(r_\infty | D_{\text{syn}}) dr_\infty. \quad (2.5) \end{aligned}$$

Thus the proportionality assumption reduces the number of variance parameters to be estimated from $k(k-1)/2$ to 1 and allows for the closed-form approximation of the integral in (2.4). As $\bar{U}_\infty = 0$, the derivation is simplified from Reiter (2005). To complete the integration, we need the distribution of $(r_\infty | D_{\text{syn}})$. Extending the scalar case in Reiter (2003), the sampling distribution of $Q^{(i)}$, the estimate of Q obtained from $D_{\text{syn}}^{(i)}$, is given by $(Q^{(i)} | Q_{\text{cen}}, B_\infty) \sim N(Q_{\text{cen}}, B_\infty)$. Under the proportionality assumption, this becomes $(Q^{(i)} | Q_{\text{cen}}, r_\infty) \sim N(Q_{\text{cen}}, r_\infty I)$. With diffuse priors and standard multivariate normal theory for sample covariance matrices, we obtain

$$(m-1) \frac{\sum_{i=1}^m (Q^{(i)} - \bar{Q}_m)(Q^{(i)} - \bar{Q}_m)'}{(m-1)r_\infty} | D_{\text{syn}} \sim \text{Wish}(m-1, I).$$

Taking the trace of each side and integrating over r_∞ in (2.5) yields a Bayesian p -value of

$$P\left(\frac{\chi_k^2}{k} \frac{k(m-1)}{\chi_{k, k(m-1)}^2} > S_c | D_{\text{syn}}\right) = P(F_{k, k(m-1)} > S_c | D_{\text{syn}}).$$

3. Missing data

The extension to missing data is straightforward. When $\bar{U}_\infty = 0$, the combining rules (Rubin 1987) for scalar estimands q simplify so that $(q | D_{\text{com}}) \sim N(\bar{q}_m, (1 + 1/m)B_m)$, where D_{com} is the set of m completed datasets. Similar to Section 2, the tests of Rubin (1987) and Li, Raghunathan and Rubin (1991) for multivariate components rely on the assumption that $B_\infty \propto \bar{U}_\infty$, and so when $\bar{U}_\infty = 0$ we derive a test under the assumption $B_\infty = r_\infty I$.

Following derivation procedures similar to that of Section 2.1, the Bayesian p -value for testing $H: Q = Q_0$ with k -dimensional Q is found to be $P(F_{k, k(m-1)} > S_q | D_{\text{com}})$ where

$$S_q = \frac{(Q_0 - \bar{Q}_m)'(Q_0 - \bar{Q}_m)}{kr_q},$$

and $r_q = (1 + 1/m) \text{tr}(B_m)/k$.

4. Simulation study

In this section, simple simulation examples illustrate the analytic validity of the proposed combining rules, first for the case of partially synthetic data, and then for the case missing data. Lastly, the robustness of the tests to the proportionality assumption is evaluated.

For a population of $N = 50,000$, $X = (X_1, \dots, X_{20})$ is drawn from a multivariate normal distribution with mean zero and covariance matrix with 1 in each diagonal element and 0.5 in each off-diagonal element. Y is drawn from a standard normal distribution. For each of 5,000 iterations, a new finite population is generated and m imputations are drawn for $m \in \{2, 5, 10\}$. The proposed hypothesis tests are conducted for $H_0: Q = Q_0$, where Q is the vector of regression coefficients, excluding the intercept, of the regression of Y on X and has dimension k , $k \in \{2, 5, 20\}$, and Q_0 is the true value of Q determined from the finite population (X, Y) . Since H_0 is true by design, H_0 should be rejected 100 $\alpha\%$ of the time, for significance level $\alpha = 0.05$.

Random sampling scenarios are also simulated for comparison purposes. At each iteration, a random sample of size $s = 50,000$ from an infinite population is generated from the distributions described above, prior to generating the m missing data and synthetic imputations. The same hypothesis $H_0: Q = Q_0$ is tested where Q_0 is the vector of true population values. The combining rules for the hypothesis tests are those of Reiter (2005) in the synthetic data case and Li *et al.* (1991) and Rubin (1987) in the missing data case.

4.1 Partially synthetic data imputations

Let Y be a confidential response variable and X be unreplaced predictors. Then Y_{syn} is generated by taking m independent draws from the posterior predictive distribution $f(Y|X)$ assuming a normal linear model, using all available data.

Table 1 gives the nominal 5% rejection rate for the proposed hypothesis test for multicomponent estimands, which are seen to be close to the significance level 0.05, and close to the random sampling results. From these results it appears that the proposed combining rules for population data have good frequentist properties. Not shown are the rejection rates when the rules from random samples (Reiter 2005) were applied to finite populations, which were observed to be quite high, typically 1, in the simulations conducted.

Table 1
Comparison of nominal 5% rejection rates for tests on partially synthetic data

	$k = 2$	$k = 5$	$k = 20$
Census data			
$m = 2$	0.048	0.065	0.052
$m = 5$	0.048	0.061	0.057
$m = 10$	0.051	0.067	0.055
Random sampling			
$m = 2$	0.067	0.062	0.060
$m = 5$	0.054	0.052	0.050
$m = 10$	0.047	0.049	0.049

4.2 Missing data

Simulations analogous to the synthetic data simulations were conducted for the missing data case. The missing values of Y are imputed from the posterior predictive distribution $f(Y_{\text{obs}} | X)$ assuming a normal linear model. Missingness is simulated to be completely at random, with $P(R_l = 1) = 0.3$, $l = 1, \dots, s$, where R is an indicator variable for missingness.

Table 2 gives the nominal 5% rejection rate for the proposed hypothesis test for multicomponent estimands, which are seen to be close to 0.05, and to the random sampling results. From these results it appears that the proposed combining rules for population data yield valid inferences.

Table 2
Comparison of nominal 5% rejection rates for tests using completed census data

	$k = 2$	$k = 5$	$k = 20$
Census data			
$m = 2$	0.052	0.061	0.053
$m = 5$	0.048	0.063	0.051
$m = 10$	0.048	0.058	0.054
Random sampling			
$m = 2$	0.061	0.056	0.053
$m = 5$	0.056	0.052	0.052
$m = 10$	0.048	0.050	0.051

4.3 Robustness

The assumption that $B_\infty \propto r_\infty I$ is striking at first glance, and is unlikely to be exactly true. In this section we evaluate the effect of strong correlations across components of Q . While moderately strong correlations were present in the previous simulations, here we increase the magnitude of the between-imputation variance, increasing the magnitude of the differences across the diagonal of B as well as the distance from zero of the off-diagonal elements of B .

These simulations are set up as before, for the finite population case, with $k = 5$ and $m = 5$. The population in each iteration is generated in the same way as before, except that we let $Y = (1, 2, 5, 10, 20, 0, \dots, 0) (X_1, X_2, \dots, X_{20})' + \eta$, $\eta \sim N(0, 100)$ and $X_2 = c \cdot X_1 + \varepsilon$, $c \in \{0.5, 1, 5\}$ and $\varepsilon \sim N(0, 1)$. Increasing values of c yields increasingly higher correlations. The large variance for η induces larger and more variable values for elements of B .

The results in Table 3 indicate that while the tests have good properties even with moderately high violations of the proportionality assumption, their performance declines with increasingly large correlations. Continuing our assumption that Q represents a vector of regression coefficients, presence of such large correlation may also be indicative of multicollinearity in the model at hand, so analysts faced with high correlation across $Q^{(i)}$ might take steps to reduce multicollinearity before applying the proposed tests. If

variables are of substantially differing magnitude, standardization to rescale them will reduce differences across Q .

Table 3
Evaluation of tests under assumption violations, $k = 5, m = 5$

	$c = 0.5$	$c = 1$	$c = 5$
Synthetic Data	0.059	0.083	0.145
Missing Data	0.051	0.083	0.136

Acknowledgements

A portion of this work was conducted while the author was a student at Duke University, supported by NSF grant ITR-0427889 and under the guidance of Jerry Reiter, whose assistance is greatly appreciated. In addition, the comments of anonymous reviewers were quite helpful.

References

Deming, W.E., and Stephan, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 213, 45-49.

Kinney, S.K., Reiter, J.P., Reznick, A.P., Miranda, J., Jarmin, R.S. and Abowd, J.M. (2011). Toward unrestricted public-use business microdata: The Longitudinal Business Database. *International Statistical Review*, 79, 3, 362-384.

Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 2, 181-188.

Reiter, J.P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131, 365-377.

Reiter, J.P., and Kinney, S.K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. Technical report, National Institute of Statistical Sciences.

Reiter, J.P., and Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.

NOTICE

Statistics Canada will be discontinuing its practice to print *Survey Methodology*. This current issue (December 2012 – volume 38 number 2) will be the last version available in print form. Please note that the electronic version of *Survey Methodology* will continue to be available free of charge on the Statistics Canada website, www.statcan.gc.ca.

The next issue is to be published in June 2013 in electronic format and will maintain our high standard of content.

You may subscribe to “My Account” on Statistics Canada website to receive email notifications when new issues of the journal are released.

CORRIGENDUM

James Chipperfield and John Preston

“Efficient bootstrap for business surveys”, vol. 33, no. 2 (December 2007), 167-172.

In Section 4.2 of this paper, under the equation

$$\text{Var}(\hat{v}_{\text{boot}}) = \text{Var}_s \left(E_* \left[\hat{v}_{\text{boot}} | s \right] \right) + E_s \left(\text{Var}_* \left[\hat{v}_{\text{boot}} | s \right] \right),$$

there are five references to the term

$$\text{Var}_s \left(E_* \left[\hat{v}_{\text{boot}} | s \right] \right).$$

To be correct, these five referenced terms should be replaced by

$$E_s \left(\text{Var}_* \left[\hat{v}_{\text{boot}} | s \right] \right).$$

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2012.

- S.R. Amer, *RTI International*
 T. Asparouhov, *Mplus*
 M. Barron, *NORC*
 W. Bell, *U.S. Census Bureau*
 E. Berg, *National Agricultural Statistical Services*
 P. Biemer, *RTI*
 I. Bilgen, *NORC*
 C. Bocci, *Statistics Canada*
 J. van den Brakel, *Statistics Netherlands*
 M. Brick, *Westat, Inc.*
 R. Bruni, *University of Rome, La Sapienza*
 C.-T. Chao, *National Cheng-Kung University, Taiwan*
 G. Chauvet, *CREST-ENSAI*
 J. Chipperfield, *Australian Bureau of Statistics*
 G. Datta, *University of Georgia*
 M. Davern, *NORC*
 T. DeWaal, *Statistics Netherlands*
 D. Dolson, *Statistics Canada*
 S. Eckman, *Institute for Employment Research, Germany*
 S. Er, *Istanbul University*
 E. Escobar, *University of Southampton*
 V. Estevao, *Statistics Canada*
 O.P. Fischer, *U.S. Census Bureau*
 J. Gambino, *Statistics Canada*
 N. Ganesh, *NORC at University of Chicago*
 T.I. Garner, *U.S. Bureau of Labor Statistics*
 J. Garrett, *Knowledge Networks, Inc.*
 C. Goga, *Université de Bourgogne*
 M. Graf, *Office fédéral de la Statistique, Suisse*
 B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
 D. Kasprzyk, *NORC at the University of Chicago*
 C. Kennedy, *Abt SRBI*
 M.G.M. Khan, *University of the South Pacific, Fiji*
 J.-K. Kim, *Iowa State University*
 P. Kott, *RTI*
 P. Lavallée, *Statistics Canada*
 F. Li, *Duke University*
 J. Li, *Westat Inc.*
 P. Lugtig, *Utrecht University*
 P. Lynn, *University of Essex*
 D. Malec, *National Center for Health Statistics*
 H. Mantel, *Statistics Canada*
 I. Molina, *Universidad Carlos III de Madrid*
 R. Münnich, *Economic and Social Statistics Dept. Univ. of Trier, Germany*
 J. Oleson, *University of Iowa*
 A.J. O'Malley, *Harvard Medical School*
 J. Opsomer, *Colorado State University*
 V. Parsons, *National Center for Health Statistics*
 D. Pfeffermann, *Hebrew University*
 F. van de Pol, *Statistics Netherlands*
 N.G.N. Prasad, *University of Alberta*
 L. Qualité, *Université de Neuchâtel*
 T. Raghunathan, *University of Michigan*
 J.N.K. Rao, *Carleton University*
 J. Reiter, *Duke University*
 L.-P. Rivest, *Université Laval*
 R. Rodriguez, *U.S. Census Bureau*
 K. Rust, *Westat, Inc.*
 E. Saleh, *Carleton University*
 F. Scheuren, *NORC*
 A. Scott, *University of Auckland*
 J. Sedransk, *Case Western Reserve University & University of Maryland*
 P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
 R. Sigman, *Westat Inc.*
 A. Singh, *NORC*
 C. Skinner, *London School of Economics*
 P.A. Smith, *Office for National Statistics*
 P.W.F. Smith, *University of Southampton*
 N. Thomas, *Pfizer*
 R. Thomas, *Carleton University*
 K.J. Thompson, *U.S. Census Bureau*
 M. Thompson, *University of Waterloo*
 Y. Tillé, *Université de Neuchâtel*
 V. Toepoel, *Tilburg University*
 M. Torabi, *University of Manitoba*
 V. Vehovar, *University of Ljubljana*
 J. Vermunt, *Tilburg School of Social and Behavioral Sciences*
 M. de Toledo Vieira, *Universidade Federal de Juiz de Fora, Brazil*
 J. Wagner, *University of Michigan*
 K. Wolter, *NORC*
 C. Wu, *University of Waterloo*
 C. Yu, *Iowa State University*
 W. Yung, *Statistics Canada*
 E. Zanutto, *National Analysts Worldwide*

Acknowledgements are also due to those who assisted during the production of the 2012 issues: Céline Ethier of Statistical Research and Innovation Division, Christine Cousineau of Household Survey Methods Division, Nick Budko and Annette Everett of Business Survey Methods Division, Anne-Marie Fleury of Operations and Integration Division, Roberto Guido, Liliane Lanoie, Darquise Pellerin, Joseph Prince, Jacqueline Luffman, Suzanne Bélair, Janice Burr, Jeff Campbell, Kathy Charbonneau and Fadi Salibi of Dissemination Division.

ANNOUNCEMENTS

Nominations Sought for the 2014 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg to recognize his contributions to survey methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium and give the 2014 Waksberg Invited Address at the Statistics Canada Symposium to be held in the autumn of 2014. The paper will be published in a future issue of *Survey Methodology* (targeted for December 2014).

The author of the 2014 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2013 to the chair of the committee, Steve Heeringa (sheering@isr.umich.edu).

Previous Waksberg Award honorees and their invited papers are:

- 2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future?". *Survey Methodology*, vol. 27, 1, 7-31.
- 2002 Wayne A. **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.
- 2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.
- 2004 Norman M. **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.
- 2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.
- 2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.
- 2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.
- 2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.
- 2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.
- 2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.
- 2011 Danny **Pfeffermann**, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". *Survey Methodology*, vol. 37, 2, 115-136.
- 2012 Lars **Lyberg**, "Survey Quality". *Survey Methodology*, vol. 38, 2, 107-130.
- 2013 Ken **Brewer**, Manuscript topic under consideration.

Members of the Waksberg Paper Selection Committee (2012-2013)

Steve Heeringa, *University of Michigan* (Chair)

Cynthia Clark, *USDA*

Louis-Paul Rivest, *Université de Laval*

J.N.K. Rao, *Carleton University*

Past Chairs:

Graham Kalton (1999 - 2001)

Chris Skinner (2001 - 2002)

David A. Binder (2002 - 2003)

J. Michael Brick (2003 - 2004)

David R. Bellhouse (2004 - 2005)

Gordon Brackstone (2005 - 2006)

Sharon Lohr (2006 - 2007)

Robert Groves (2007 - 2008)

Leyla Mojadjer (2008 - 2009)

Daniel Kasprzyk (2009 - 2010)

Elizabeth A. Martin (2010 - 2011)

Mary E. Thompson (2011 - 2012)

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 28, No. 2, 2012

Collecting Survey Data During Armed Conflict William G. Axinn, Dirgha Ghimire, Nathalie E. Williams	153
Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures Jeffrey A. Groen.....	173
Management Challenges of the 2010 U.S. Census Daniel H. Weinberg	199
Response Rates in Business Surveys: Going Beyond the Usual Performance Measure Katherine Jenny Thompson, Broderick E. Oliver.....	221
Calibration Inspired by Semiparametric Regression as a Treatment for Nonresponse Giorgio E. Montanari, M. Giovanna Ranalli	239
Strategy for Modelling Nonrandom Missing Data Mechanisms in Observational Studies Using Bayesian Methods Alexina Mason, Sylvia Richardson, Ian Plewis, Nicky Best.....	279
Book Reviews.....	303

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 28, No. 3, 2012

Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics Roderick J. Little.....	309
Discussion	
Jean-François Beaumont.....	335
Philippe Brion	341
Alan H. Dorfman	349
Risto Lehtonen.....	353
Paul A. Smith.....	359
Michael P. Cohen.....	363
Rejoinder	
Roderick J. Little.....	367
Improving RDD Cell Phone Samples. Evaluation of Different Pre-call Validation Methods Tanja Kunz, Marek Fuchs	373
Mutual Information as a Measure of Inter-coder Agreement Ben Klemens.....	395
The Organization of Information in a Statistical Office Tjalling Gelsema.....	413
Unit Root Properties of Seasonal Adjustment and Related Filters William R. Bell	441
Book Review	463
In Other Journals	469

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Volume 40, No. 2, June/juin 2012

Hui Song, Yingwei Peng and Dongsheng Tu A new approach for joint modelling of longitudinal measurements and survival times with a cure fraction	207
Georgios Papageorgiou Restricted maximum likelihood estimation of joint mean-covariance models.....	225
Karelyn A. Davis, Chul G. Park and Sanjoy K. Sinha Testing for generalized linear mixed models with cluster correlated data under linear inequality constraints.....	243
David Haziza and Frédéric Picard Doubly robust point and variance estimation in the presence of imputed survey data.....	259
Jieli Ding, Yanyan Liu, David B. Peden, Steven R. Kleeberger and Haibo Zhou Regression analysis for a summed missing data problem under an outcome-dependent sampling scheme	282
Hannes Kazianka and Jürgen Pilz Objective Bayesian analysis of spatial data with uncertain nugget and range parameters	304
Tingting Zhang and Jun S. Liu Nonparametric hierarchical Bayes analysis of binomial data via Bernstein polynomial priors	328
Kei Hirose and Sadanori Konishi Variable selection via the weighted group lasso for factor analysis models	345
Zhibiao Zhao and Weixin Yao Sequential design for nonparametric inference	362
José R. Berrendero, Antonio Cuevas and Beatriz Pateiro-López Testing uniformity for the case of a planar unknown support.....	378
Acknowledgement of referees' services/Remerciements aux membres des jurys.....	396

Volume 40, No. 3, September/septembre 2012

Yulia R. Gel and Bei Chen Robust Lagrange multiplier test for detecting ARCH/GARCH effect using permutation and bootstrap.....	405
Florian Ketterer and Hajo Holzmann Testing for intercept-scale switch in linear autoregression	427
Pierre Duchesne, Kilani Ghoudi and Bruno Rémillard On testing for independence between the innovations of several time series	447
Ivan Kojadinovic and Jun Yan Goodness-of-fit testing based on a weighted bootstrap: A fast large-sample alternative to the parametric bootstrap	480
Ramon Oller and Guadalupe Gómez A generalized Fleming and Harrington's class of tests for interval-censored data	501
Carlotta Ching Ting Fok, James O. Ramsay, Michal Abrahamowicz and Paul Fortin A functional marked point process model for lupus data	517
Grace Y. Yi and Jerald F. Lawless Likelihood-based and marginal inference methods for recurrent event data with covariate measurement error	530
Hongjian Zhu and Feifang Hu Interim analysis of clinical trials based on urn models	550
Zhong Guan, Jing Qin and Biao Zhang Information borrowing methods for covariate-adjusted ROC curve	569
Jiming Jiang and Thuan Nguyen Small area estimation via heteroscedastic nested-error regression.....	588
Jae Kwang Kim and Minki Hong Imputation for statistical inference with coarse data	604

GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

1. Layout

- 1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The documents should be divided into numbered sections with suitable verbal titles.
- 1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de finaliser votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférentiellement Word. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1.	Présentation	<p>Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.</p> <p>Les textes doivent être divisés en sections numérotées portant des titres appropriés.</p> <p>Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.</p> <p>Les remerciements doivent paraître à la fin du texte.</p> <p>Toute annexe doit suivre les remerciements mais précéder la bibliographie.</p>
2.	Résumé	<p>Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.</p>
3.	Rédaction	<p>Éviter les notes au bas des pages, les abréviations et les sigles.</p> <p>Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc.</p> <p>Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.</p> <p>Écrire les fractions dans le texte à l'aide d'une barre oblique.</p> <p>Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0 ; l, 1).</p> <p>Les caractères italiques sont utilisés pour faire ressortir des mots.</p>
4.	Figures et tableaux	<p>Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).</p>
5.	Bibliographie	<p>Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).</p> <p>La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.</p>
6.	Communications brèves	<p>Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.</p>

Volume 40, No. 3, September/septembre 2012

Yulia R. Gel and Bei Chen	405
Robust Lagrange multiplier test for detecting ARCH/GARCH effect using permutation and bootstrap.....	
Florian Ketterer and Hajo Holzmann	427
Testing for intercept-scale switch in linear autoregression	
Pierre Duchesne, Kilani Ghoudi and Bruno Régnier	447
On testing for independence between the innovations of several time series	
Ivan Kojadinovic and Jun Yan	480
Goodness-of-fit testing based on a weighted bootstrap: A fast large-sample alternative to the parametric bootstrap	
Ramon Oller and Guadalupe Gómez	501
A generalized Fleming and Harrington's class of tests for interval-censored data	
Carlotia Ching Ting Fok, James O. Ramsay, Michal Abrahamowicz and Paul Fortin	517
A functional marked point process model for lupus data	
Grace Y. Yi and Gerald F. Lawless	530
Likelihood-based and marginal inference methods for recurrent event data with covariate measurement error	
Hongjian Zhu and Feifang Hu	550
Interim analysis of clinical trials based on urn models	
Zhong Guan, Jing Qin and Biao Zhang	569
Information borrowing methods for covariate-adjusted ROC curve	
Jinming Jiang and Thuan Nguyen	588
Small area estimation via heteroscedastic nested-error regression.....	
Jae Kwang Kim and Minki Hong	604
Imputation for statistical inference with coarse data	

Volume 40, No. 2, June/juin 2012

Hui Song, Yingwei Peng and Dongsheng Tu A new approach for joint modelling of longitudinal measurements and survival times with a cure fraction.....	207
Georgios Papageorgiou Restricted maximum likelihood estimation of joint mean-covariance models.....	225
Karelyn A. Davis, Chul G. Park and Sanjoy K. Sinha Testing for generalized linear mixed models with cluster correlated data under linear inequality constraints.....	243
David Haziza and Frédéric Picard Doubly robust point and variance estimation in the presence of imputed survey data.....	259
Jieli Ding, Yanyan Liu, David B. Peden, Steven R. Kleiberg and Haiibo Zhou Regression analysis for a summed missing data problem under an outcome-dependent sampling scheme.....	282
Hannes Kazianka and Jürgen Pilz Objective Bayesian analysis of spatial data with uncertain nugget and range parameters.....	304
Tingting Zhang and Jun S. Liu Nonparametric hierarchical Bayes analysis of binomial data via Bernstein polynomial priors.....	328
Kei Hirose and Sadanori Konishi Variable selection via the weighted group lasso for factor analysis models.....	345
Zhibiao Zhao and Weixin Yao Sequential design for nonparametric inference.....	362
José R. Berrendero, Antonio Cuevas and Beatriz Pateiro-López Testing uniformity for the case of a planar unknown support.....	378
Acknowledgement of referees' services/Remerciements aux membres des jurys.....	396

JOURNAL OF OFFICIAL STATISTICS
An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 28, No. 3, 2012

Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics	309
Roderick J. Little	
Discussion	
Jean-François Beaumont	335
Philippe Brion	341
Alan H. Dorfman	349
Risto Lehtonen	353
Paul A. Smith	359
Michael P. Cohen	363
Rejoinder	
Roderick J. Little	367
Improving RDD Cell Phone Samples. Evaluation of Different Pre-call Validation Methods	
Tanja Kunz, Marek Fuchs	373
Mutual Information as a Measure of Interdecoder Agreement	
Ben Klemens	395
The Organization of Information in a Statistical Office	
Tjalling Gelsma	413
Unit Root Properties of Seasonal Adjustment and Related Filters	
William R. Bell	441
Book Review	463
In Other Journals	469

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 28, No. 2, 2012

Collecting Survey Data During Armed Conflict	William G. Axinn, Dirgha Ghimire, Nathalie E. Williams	153
Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures	Jeffrey A. Groen	173
Management Challenges of the 2010 U.S. Census	Daniel H. Weinberg	199
Response Rates in Business Surveys: Going Beyond the Usual Performance Measure	Katharine Jenny Thompson, Broderick E. Oliver	221
Calibration Inspired by Semiparametric Regression as a Treatment for Nonresponse	Giorgio E. Montanari, M. Giovanna Ranalli	239
Strategy for Modelling Nonrandom Missing Data Mechanisms in Observational Studies Using Bayesian Methods	Alexina Mason, Sylvia Richardson, Ian Plewis, Nicky Best	279
Book Reviews		303

All inquiries about submissions and subscriptions should be directed to jos@sch.se

Membres du comité de sélection de l'article Waksberg (2012-2013)

Steve Heeringa, *University of Michigan* (Président)
 Cynthia Clark, *USDA*
 Louis-Paul Rivest, *Université de Laval*
 J.N.K. Rao, *Carleton University*

Présidents précédents :

Graham Kalton (1999 - 2001)
 Chris Skinner (2001 - 2002)
 David A. Binder (2002 - 2003)
 J. Michael Brick (2003 - 2004)
 David R. Bellhouse (2004 - 2005)
 Gordon Brackstone (2005 - 2006)
 Sharon Lohr (2006 - 2007)
 Robert Groves (2007 - 2008)
 Leyla Mojađer (2008 - 2009)
 Daniel Kasprzyk (2009 - 2010)
 Elizabeth A. Martin (2010 - 2011)
 Mary E. Thompson (2011 - 2012)

ANNONCES

Demande de candidatures pour le prix Waksberg 2014

La revue *Techniques d'enquête* a mis sur pied une série annuelle de communications sollicitées en l'honneur de Joseph Waksberg, en reconnaissance des contributions qu'il a faites à la méthodologie d'enquête. Chaque année, un éminent statisticien d'enquête est choisi pour rédiger un article où il examine l'évolution et l'état actuel d'un thème important du domaine de la méthodologie d'enquête. L'article reflète le mélange de théorie et de pratique caractéristique des travaux de Joe Waksberg.

Le lauréat du prix Waksberg recevra une prime en argent et présentera la communication sollicitée Waksberg 2014 au Symposium de Statistique Canada qui se tiendra à l'automne de 2014. L'article paraîtra dans un numéro de *Techniques d'enquête* (publication prévue pour décembre 2014).

L'auteur de l'article Waksberg 2014 sera choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*. Les candidatures ou les suggestions de thèmes doivent être envoyées avant le 28 février 2013 au président du comité, Steve Heeringa (sheering@isr.umich.edu).

Les gagnants et articles précédents du prix Waksberg sont

- 2001 Gad **Nathan**, « Méthodes de téléenquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir », *Techniques d'enquête*, vol. 27, 1, 7-34.
- 2002 Wayne A. **Faller**, « Estimation par régression appliquée à l'échantillonnage », *Techniques d'enquête*, vol. 28, 1, 5-25.
- 2003 David **Holt**, « Enjeux méthodologiques de l'élaboration et de l'utilisation d'indicateurs statistiques pour des fins de comparaisons internationales », *Techniques d'enquête*, vol. 29, 1, 5-19.
- 2004 Norman M. **Bradburn**, « Comprendre le processus de question et réponse », *Techniques d'enquête*, vol. 30, 1, 5-16.
- 2005 J.N.K. **Rao**, « Evaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage », *Techniques d'enquête*, vol. 31, 2, 127-151.
- 2006 Alastair **Scott**, « Etudes cas-témoins basées sur la population », *Techniques d'enquête*, vol. 32, 2, 137-147.
- 2007 Carl-Erik **Særdal**, « La méthode de calage dans la théorie et la pratique des enquêtes », *Techniques d'enquête*, vol. 33, 2, 113-135.
- 2008 Mary E. **Thompson**, « Enquêtes internationales : motifs et méthodologies », *Techniques d'enquête*, vol. 34, 2, 145-157.
- 2009 Graham **Kalton**, « Méthodes de suréchantillonnage des sous-populations rares dans les enquêtes sociales », *Techniques d'enquête*, vol. 35, 2, 133-152.
- 2010 Ivan P. **Fellegi**, « L'organisation de la méthodologie statistique et de la recherche méthodologique dans les bureaux nationaux de la statistique », *Techniques d'enquête*, vol. 36, 2, 131-139.
- 2011 Danny **Pfeffermann**, « Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ? Comment le résoudre ? », *Techniques d'enquête*, vol. 37, 2, 123-146.
- 2012 Lars **Lyberg**, « La qualité des enquêtes », *Techniques d'enquête*, vol. 38, 2, 115-142.
- 2013 Ken **Brewer**, Sujet de l'article à l'étude.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2012.

S.R. Amer, *RTI International*
T. Asparouhov, *Mplus*
M. Barron, *NORC*
W. Bell, *U.S. Census Bureau*
E. Berg, *National Agricultural Statistical Services*
P. Biemer, *RTI*
I. Bilgen, *NORC*
C. Bocci, *Statistics Canada*
J. van den Brakel, *Statistics Netherlands*
M. Brick, *Westat, Inc.*
R. Bruni, *University of Rome, La Sapienza*
C.-T. Chao, *National Cheng-Kung University, Taiwan*
G. Chauver, *CREST-ENSAI*
J. Chipperfield, *Australian Bureau of Statistics*
G. Datta, *University of Georgia*
M. Davern, *NORC*
T. DeWaal, *Statistics Netherlands*
D. Doison, *Statistics Canada*
S. Eckman, *Institute for Employment Research, Germany*
S. Er, *Instanbul University*
E. Escobar, *University of Southampton*
O.P. Fischer, *U.S. Census Bureau*
J. Gambino, *Statistics Canada*
N. Ganes, *NORC at University of Chicago*
T.J. Garner, *U.S. Bureau of Labor Statistics*
J. Garrett, *Knowledge Networks, Inc.*
C. Goga, *Université de Bourgogne*
M. Graf, *Office fédéral de la Statistique, Suisse*
B. Hülliger, *University of Applied Sciences Northwestern Switzerland*
D. Kasprzyk, *NORC at the University of Chicago*
C. Kennedy, *Abi SRBI*
M.G.M. Khan, *University of the South Pacific, Fiji*
J.-K. Kim, *Iowa State University*
P. Kott, *RTI*
P. Lavalée, *Statistics Canada*
F. Li, *Duke University*
J. Li, *Westat Inc.*
P. Lueftig, *Ulrich University*
P. Lynn, *University of Essex*
D. Malec, *National Center for Health Statistics*
W. Yung, *Statistique Canada*
E. Zanutto, *National Analysis Worldwide*
I. Molina, *Universidad Carlos III de Madrid*
R. Münich, *Economic and Social Statistics Dept. Univ. of Trier, Germany*
J. Oleson, *University of Iowa*
A.J. O'Malley, *Harvard Medical School*
J. Opsomer, *Colorado State University*
V. Parsons, *National Center for Health Statistics*
D. Pfeffermann, *Hebrew University*
F. van de Pol, *Statistics Netherlands*
N.G.N. Prasad, *University of Alberta*
L. Quailé, *Université de Neuchâtel*
T. Raghunathan, *University of Michigan*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
R. Rodriguez, *U.S. Census Bureau*
K. Rust, *Westat, Inc.*
E. Saleh, *Carleton University*
F. Scheuren, *NORC*
A. Scott, *University of Auckland*
J. Sedransk, *Case Western Reserve University & University of Maryland*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
R. Sigman, *Westat Inc.*
A. Singh, *NORC*
C. Skinner, *London School of Economics*
P.A. Smith, *Office for National Statistics*
P.W.F. Smith, *University of Southampton*
N. Thomas, *Pfizer*
R. Thomas, *Carleton University*
K.J. Thompson, *U.S. Census Bureau*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
V. Toepoel, *Tilburg University*
M. Torabi, *University of Manitoba*
V. Vohovar, *University of Ljubljana*
J. Vermunt, *Tilburg School of Social and Behavioral Sciences*
M. de Toledo Vieira, *Universidade Federal de Juiz de Fora, Brazil*
J. Wagner, *University of Michigan*
K. Wolter, *NORC*
C. Wu, *University of Waterloo*
C. Yu, *Iowa State University*
W. Yung, *Statistique Canada*
E. Zanutto, *National Analysis Worldwide*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2012 : Céline Ethier de la Division de la recherche et de l'innovation en statistique, Christine Cousineau de la Division des méthodes d'enquêtes auprès des ménages, Nick Budko et Annette Everett de la Division des méthodes d'enquêtes auprès des centres de données, Joseph Prince, Jacqueline Luffman, Suzanne Bélair, Janice Burt, Jeff Campbell, Kathy Charbonneau et Fadi Salibi de la Division de la diffusion.

CORRIGENDUM

James Chipperfield et John Preston,
« Bootstrap efficace pour les enquêtes-entrepises », vol. 33, n° 2 (Décembre 2007), 187-193.
À la section 4.2 de cet article, sous l'équation

$$\text{Var}(\hat{\psi}^{\text{boot}}) = \text{Var}_s(E_s[\hat{\psi}^{\text{boot}}|s]) + E_s(\text{Var}_s[\hat{\psi}^{\text{boot}}|s]),$$

l'exposé contient cinq mentions du terme

$$\text{Var}_s(E_s[\hat{\psi}^{\text{boot}}|s]).$$

Pour que l'exposé soit correct, ces cinq mentions doivent être remplacées par

$$E_s(\text{Var}_s[\hat{\psi}^{\text{boot}}|s]).$$

AVERTISSEMENT

Statistique Canada cessera de publier une version imprimée de la revue *Techniques d'enquête*. Ce plus récent numéro (décembre 2012 – volume 38, numéro 2) sera le dernier disponible en version imprimée. Veuillez noter que la version électronique de *Techniques d'enquête* demeurera disponible gratuitement sur le site internet de Statistique Canada, www.statcan.gc.ca.

Notre prochain numéro sera diffusé en juin 2013 en version électronique selon nos mêmes normes rigoureuses quant au contenu.

Vous pouvez vous inscrire sous « Mon compte » sur le site internet de Statistique Canada pour recevoir un avis par courriel lors de la publication des prochains numéros de la revue.

Tableau 2
Comparaison des taux de rejet au niveau nominal de 5 % pour les tests sur des données de recensement complètes

Données de recensement			Échantillonnage aléatoire		
$k = 2$	$k = 5$	$k = 20$	$m = 2$	$m = 5$	$m = 10$
0,052	0,061	0,053	0,048	0,052	0,053
0,048	0,063	0,051	0,056	0,052	0,051
0,048	0,058	0,054	0,061	0,056	0,051

4.3 Robustesse

L'hypothèse que $B_{\infty} \propto r_{\infty}^2 I$ est étonnante à première vue, et il est peu probable qu'elle soit exactement vraie. À la présente section, nous évaluons l'effet de fortes corrélations entre les composantes de \tilde{Q} . Même si des corrélations moyennement fortes étaient présentes dans les simulations précédentes, ici, nous augmentons la grandeur de la variance entre imputations, ce qui accroît les écarts le long de la diagonale de B ainsi que la distance par rapport à zéro des éléments hors diagonale de B .

Les simulations sont configurées comme précédemment, pour le cas d'une population finie, avec $k = 5$ et $m = 5$. À chaque itération, la population est générée de la même façon qu'auparavant, excepté que nous prenons $Y = (1, 2, 5, 10, 20, 0, \dots, 0)$ (X_1, X_2, \dots, X_{20}) + η , $\eta \sim N(0, 100)$ et $X_2 = c \cdot X_1 + e$, $e \in \{1/2, 1, 5\}$ et $e \sim N(0, 1)$. Des valeurs croissantes de c donnent des corrélations de plus en plus fortes. La grande variance de η induit des valeurs grandes et plus variables des éléments de B .

Le tableau 3 montre que, bien que les tests aient de bonnes propriétés même sous des violations moyennement fortes de l'hypothèse de proportionnalité, leur performance diminue à mesure qu'augmente la force des corrélations. En maintenant l'hypothèse que \tilde{Q} représente un vecteur de coefficients de régression, l'existence d'une corrélation aussi forte peut également être un signe de multicollinéarité dans le modèle utilisé, de sorte que les analystes en présence d'une forte corrélation entre les $\tilde{Q}^{(i)}$ pourraient vouloir prendre des mesures en vue de réduire la multicollinéarité avant d'appliquer les tests proposés. Si les variables sont de grandeur très différente, une normalisation en vue de les rééchantillonner réduira les écarts entre les \tilde{Q} .

Bibliographie

Tableau 3
Évaluation des tests sous violations d'hypothèses, $k = 5$, $m = 5$

$c = 1$		$c = 1/2$		$c = 5$	
Données synthétiques	0,083	0,059	0,083	0,145	0,136
Données manquantes	0,051	0,051	0,083	0,136	0,136

Une partie des travaux susmentionnés, financée par la bourse ITR-0427889 de la NSF, a été effectuée pendant que l'auteur était étudiant à la Duke University, sous la supervision de Jerry Reiter, dont l'aide a été fort appréciée. En outre, les commentaires d'examineurs anonymes ont été très utiles.

Remerciements

Deming, W.E., et Stephan, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 213, 45-49.

Kinney, S.K., Reiter, J.P., Rennek, A.P., Miranda, J., Jarmain, R.S., et Abowd, J.M. (2011). Toward unrestricted public-use business microdata: The Longitudinal Business Database. *Revue Internationale de Statistique*, 79, 3, 362-384.

Li, K.H., Raghunathan, T.E. et Rubin, D.B. (1991). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Reiter, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 2, 203-211.

Reiter, J.P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131, 365-377.

Reiter, J.P., et Kinney, S.K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. Rapport technique, National Institute of Statistical Sciences.

Reiter, J.P., et Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.

4.1 Imputation de données partiellement synthétiques

Soit Y la variable réponse confidentielle et X , les variables explicatives non remplacées. Alors, X^{syn} est générée en effectuant m tirages indépendants à partir de la loi prédictrive a posteriori $f(X | X)$ sous l'hypothèse d'un modèle linéaire normal, en utilisant toutes les données disponibles. Le tableau 1 donne les taux de rejet au niveau nominal de 5 % pour le test d'hypothèse proposé pour des quantités à estimer à composantes multiples, et montre qu'ils sont proches du seuil de signification de 0,05, ainsi que de ceux obtenus sous échantillonnage aléatoire. Ces résultats semblent indiquer que les règles de combinaison proposées pour les données de population ont de bonnes propriétés fréquentistes. Les taux de rejet obtenus en appliquant les règles établies pour les échantillons aléatoires (Reiter 2005) à des populations finies, lesquels étaient relativement élevés, habituellement égaux à 1, dans les simulations exécutées ne sont pas présentes.

Tableau 1
Comparaison des taux de rejet au niveau nominal de 5 % pour les tests sur des données partiellement synthétiques

Données de recensement		Échantillonnage aléatoire	
$m = 2$	0,048	$m = 2$	0,067
$m = 5$	0,048	$m = 5$	0,054
$m = 10$	0,051	$m = 10$	0,047
$k = 2$	0,065	$k = 2$	0,062
$k = 5$	0,061	$k = 5$	0,052
$k = 20$	0,057	$k = 20$	0,049

Des simulations analogues à celles exécutées sur les données synthétiques ont été effectuées dans le cas des données manquantes. Les valeurs manquantes de X ont été imputées à partir de la loi prédictrive a posteriori $f(X^{obs} | X)$ sous l'hypothèse d'un modèle linéaire normal. Les données manquantes ont été simulées comme si elles manquaient entièrement au hasard, avec $P(R_l = 1) = 0,3$, $l = 1, \dots, s$, où R est une variable indicatrice de l'absence de données.

Le tableau 2 donne les taux de rejet au niveau nominal de 5 % pour le test d'hypothèse proposé pour des grandeurs à composantes multiples, et montre qu'ils sont proches de 0,05, ainsi que des valeurs obtenues sous échantillonnage aléatoire. Ces résultats semblent indiquer que les règles de combinaison proposées pour des données de population donnent des inférences valides.

4.2 Données manquantes

façon que $(q | D^{com}) \sim N(\bar{q}^m, (1 + 1/m)B^m)$, où D^{com} est le jeu de m ensembles de données complètes. Comme à la section 2, les tests de Rubin (1987) et de Li, Ragunathan et Rubin (1991) pour les composantes multivariées dépendent de l'hypothèse que $B_\infty \propto U_\infty$, et donc, quand $U_\infty = 0$, nous déterminons un test sous l'hypothèse que $B_\infty = r_\infty I$.

En suivant des méthodes de calcul semblables à celles de la section 2.1, la valeur p bayésienne obtenue pour tester l'hypothèse $H : \bar{Q} = \bar{Q}_0$ avec \bar{Q} de dimension k est $P(F_{k, s(m-1)} > S_g^p | D^{com})$, où

$$S_g^p = \frac{(\bar{Q}_0 - \bar{Q}_m)(\bar{Q}_0 - \bar{Q}_m)^{k_g}}{k_g},$$
$$\text{et } r_g^p = (1 + 1/m)tr(B^m)/k.$$

4. Étude par simulation

À la présente section, des exemples de simulations simples illustrent la validité analytique des règles de combinaison proposées, d'abord pour le cas de données partiellement synthétiques, puis pour celui de données manquantes. Enfin, la robustesse des tests à l'hypothèse de proportionnalité est évaluée.

Pour une population de $N = 50\,000$, $X = (X^1, \dots, X^{20})$ est tiré d'une loi normale multivariée de moyenne nulle et de matrice de covariance où chaque élément diagonal est égal à 1 et chaque élément hors diagonale est égal à 0,5. X est tiré d'une loi normale centrée réduite. Pour chacune des 5 000 itérations, une nouvelle population finie est générée et m imputations sont tirées pour $m \in \{2, 5, 10\}$. Les tests d'hypothèse proposés sont effectués pour $H_0 : \bar{Q} = \bar{Q}_0$, où \bar{Q} est le vecteur des coefficients de régression, à l'exclusion de l'ordonnée à l'origine, de la régression de X sur X , et est de dimension k , $k \in \{2, 5, 20\}$, et \bar{Q}_0 est la valeur réelle de \bar{Q} déterminée à partir de la population finie (X, Y) . Puisque l'hypothèse nulle H_0 est vraie par conception, elle devrait être rejetée dans 100 % des cas, pour un seuil de signification de $\alpha = 0,05$.

Des scénarios d'échantillonnage aléatoire sont également simulés aux fins de comparaison. À chaque itération, un échantillon aléatoire de taille $s = 50\,000$ d'une population infinie est généré à partir de la distribution décrite plus haut, avant de générer les m données manquantes et imputations synthétiques. La même hypothèse $H_0 : \bar{Q} = \bar{Q}_0$ est testée, où \bar{Q}_0 est le vecteur de valeurs de population réelles. Les règles de combinaison pour les tests d'hypothèse sont celles de Reiter (2005) dans le cas des données synthétiques et celles de Li et coll. (1991) et de Rubin (1987) dans le cas des données manquantes.

le cas d'une population finie. Un analyste qui a accès à D_{syn} mais non à D_{rec} peut obtenir des inférences valides pour une grandeur scalaire ou vectorielle \tilde{Q} en utilisant les quantités suivantes :

partiellement synthétiques générés à l'aide de paramètres de modèle d'imputation fixes ainsi qu'aléatoires. Les simulations (non présentées) confirment qu'il en est de même dans

$$(2.1) \quad {}_{(t)}\bar{\sigma} \sum_{m=1}^{l-1} \frac{m}{1} = {}^m\bar{\sigma}$$

$$(2.2) \quad \sum_{m=1}^{l-1} \frac{m}{l} = \frac{l-1}{2}$$

$$(2.3) \quad (\tilde{\partial}^m - {}_{(t)}\tilde{\partial})(\tilde{\partial}^m - {}_{(t)}\tilde{\partial}) \sum_m^{l=1} \frac{1-m}{1} = B_m$$

où $\hat{Q}_{(i)}$, $i = 1, \dots, m$ est l'estimation ponctuelle de \bar{Q} obtenue à partir de $D_{(i)}^{\text{syn}}$, $U_{(i)}$ est la variance estimée de \bar{Q} , et B^m est la variance d'échantillon des $\hat{Q}_{(i)}$, $i = 1, \dots, m$.

B^m est la variance d'échantillon des $\widehat{Q}_{(i)}^m, i = 1, \dots, m$.

En l'absence de variance d'échantillonnage, les règles de

combinaison pour la grandeur scalaire \tilde{Q} établie par Reiter (2003) peuvent être appliquées comme cas particulier où $\underline{U}^m = 0$. La simplification résultante signifie que les ap-

proximations de Reiter (2003) ne sont pas nécessaires et que la loi a posteriori exacte sous la théorie normale multivariée

est $(\hat{\theta} | \hat{\theta}^{\text{syn}}) \sim \mathcal{N}(m, m^{-1})$. Cependant, pour un vecteur \hat{Q} , le test d'hypothèse de Reiter (2005) dépend de l'hypothèse que $B_{\hat{Q}}$ est proportionnelle à $\hat{U}_{\hat{Q}}$, c'est-à-dire

que la proportion de l'information remplacée par des imputations est la même pour toutes les composantes de \hat{Q} ,

de sorte qu'une hypothèse différente est nécessaire pour le cas où $\underline{U}^\infty = 0$.

2.1 Test multivarié proposé

A la présente section, un test de remplacement est établi en se fondant sur l'hypothèse plus forte que $B^\infty = r^\infty I$, pour une quantité scalaire r^∞ et une matrice identité I de

dimension k . Autrement dit, la variance entre imputations est constante pour toutes les composantes de \hat{Q} , et B_{∞} est supposée diagonale. Tant dans le test de Reiter (2005) que dans le test proposé, on calcule la moyenne sur l'ensemble

des composantes de la variance, de sorte que le test est moyennement robuste à cette hypothèse : cependant, la validité de la randomisation diminue quand les estimations de \bar{Q}_i , $i = 1, \dots, m$, sont fortement corrélées. Cet aspect est évalué à l'aide de simulations à la section 4.3. Des tests comparables fondés sur l'hypothèse que $B \propto \bar{U}^\infty$ perdent, on le sait, de la puissance quand l'hypothèse n'est pas

Le test proposé pour l'hypothèse $H^0: \tilde{Q} = \tilde{Q}_0$ est effectué en supposant que la statistique de test

Kinney : imputation multiple dans le cas de données de recensement

$$\frac{{}^2\lambda q}{({}^u\bar{\partial} - {}^0\bar{\partial})({}^u\bar{\partial} - {}^0\bar{\partial})} = {}^2S$$

suit une loi $F_{k, k(m-1)}^c$, où $r^c = 1 / m \operatorname{tr}(B_m) / k$. Sous l'hypothèse que $B_\infty = r^\infty I$, la valeur p bayésienne est donnée par

$$\int P(\chi^k_2 > (\bar{\mathcal{O}} - {}^0\bar{\mathcal{O}})'T^{-1}_\infty(\bar{\mathcal{O}} - {}^0\bar{\mathcal{O}}) \mid D^{\text{sym}}, B^\infty)$$

$$\left(\int P \chi_2 < \frac{m^\infty}{(\bar{\mathcal{O}}_0 - I)'(\bar{\mathcal{O}}_0 - I)} \mid D^{\text{syn}}, r^\infty \right)$$

$$P(r_\infty | D_{\text{sym}}) dr_\infty = \int d\left(\frac{\chi_k}{\chi_\ell}\right) \cdot \frac{m_r}{r_\infty} \cdot S^c |D_{\text{sym}}|_{r_\infty}$$

$$(2.5) \quad P(r_\infty | D_{\text{syn}}^\infty) dr_\infty.$$

Donc, l'hypothèse de proportionnalité réduit le nombre de paramètres de variance à estimer qui passe de $k(k-1)/2$ à 1, et permet une approximation analytique de l'intégrale en (2.4). Comme $\int_0^\infty \omega = 0$, les calculs sont plus

simples que dans Reiter (2005). Pour achever l'intégration, nous avons besoin de la distribution de $(r \mid D^{\text{syn}})$. En

et pendant le cas scalaire donne Reiter (2003), la distribution d'échantillonnage de $\hat{Q}_{(t)}$, l'estimation de \hat{Q} obtenue à partir de $D_{(t)}$ est donnée par $(Q_{(t)} | Q, B)$.

pression devient $(O_{(t)}^{\text{rec}}, r^{\infty}) \sim N(O_{(t)}^{\text{rec}}, r^{\infty})$. En utili-

sant des lois a priori diffuses et en appliquant la théorie normale multivariée classique pour les matrices de co-

3. Données manquantes

$$({}^{\cdot}D^{\text{sym}}|S^c|D^F)^{k,k(m-1)} = \left({}^{\cdot}D^{\text{sym}}|S^c|D^F \right)^{k,k(m-1)} = \left({}^{\cdot}D^{\text{sym}}|S^c|D^F \right)^{k,k(m-1)}$$

En prenant la trace de chaque membre de l'équation et en intégrant sur r_∞ dans (2.5), nous obtenons une valeur D bayésienne de

$$D^{\text{syn}}_{\text{syn}} \sim \frac{(1-m)^{\infty} r}{\sum_m (1-m)^{\underline{\omega}_{(t)}} \underline{\omega}_{(t)}}.$$

Imputation multiple dans le cas de données de recensement

Sakartar K. Kinney¹

Résumé

L'un des avantages de l'imputation multiple est qu'elle permet aux utilisateurs des données de faire des inférences valides en appliquant des méthodes classiques avec des règles de combinaison simples. Toutefois, les règles de combinaison établies pour les tests d'hypothèse multivariés échouent quand l'erreur d'échantillonnage est nulle. Le présent article propose des tests modifiés utilisables dans les analyses en population finie de données de recensement comportant de multiples imputations pour contrôler la divulgation et remplacer des données manquantes, et donne une évaluation de leurs propriétés fréquentistes par simulation.

Mots clés : Populations finies ; données manquantes ; test de signification ; données synthétiques.

1. Introduction

L'imputation multiple a été proposée au départ pour traiter la non-réponse dans les grandes enquêtes complexes (Rubin 1987). Depuis, plusieurs autres usages ont été

suggérés, dont le contrôle de la divulgation statistique et la correction de l'erreur de mesure. L'un des attraits de l'imputation multiple tient au fait que l'on peut appliquer des méthodes classiques à chaque ensemble de données imputé, puis utiliser des règles de combinaison, qui varient selon l'application. Voir Reiter et Raghunathan (2007) pour une revue détaillée des différentes règles et applications. Les règles existantes de combinaison sous imputation multiple ont été établies pour des échantillons aléatoires et des modèles de superpopulation (Deming et Stephan 1941). Dans les analyses de données de recensement en population finie, où la variance d'échantillonnage est nulle, les règles de combinaison applicables aux paramètres à estimer univariés peuvent encore l'être en tant que cas particulier ; par contre, les tests d'hypothèse échouent pour les paramètres multivariés.

Motivé par l'utilisation de l'imputation multiple pour produire des données partiellement synthétiques (Rubin 1993 ; Little 1993) pour la base de données longitudinales sur les entreprises du U.S. Census Bureau (Kinney, Reiter, Reznick, Miranda, Jarmín et Abowd 2011), c'est-à-dire un recensement économique, le présent article décrit l'élaboration d'un test multivarié pour populations finies applicable à des données partiellement synthétiques et son extension à l'imputation de données manquantes. Les extensions à d'autres applications d'imputation multiple devraient être simples.

La présentation de l'article est la suivante. La section 2 décrit le cas de données partiellement synthétiques et la section 3, l'extension aux données manquantes. Enfin, la section 4 décrit les simulations en vue d'évaluer les règles

de combinaison pour le cas des données manquantes ainsi que celui des données partiellement synthétiques.

2. Données partiellement synthétiques

Pour créer des ensembles de données partiellement synthétiques, on remplace certaines valeurs des données confidentielles par m tirages indépendants à partir de leur loi prédictrive à posteriori. Pour une population finie de taille N , soit $Z_j = 1, j = 1, \dots, N$ indiquant que l'unité j a été sélectionnée pour le remplacement par imputation de n imputations et X la matrice de variables dont les valeurs ne seront pas remplacées. Représentons par $D_{\text{rec}} = (X, Y)$ un recensement des N unités contenant des données confidentielles et supposons que toutes les unités sont entièrement observées, c'est-à-dire qu'il n'existe aucune valeur manquante. Soit $Y_{\text{rec}}^{(i)}, i = 1, \dots, m$ la i^{e} imputation de X , et soit $D_{\text{syn}}^{(i)} = (X, Y_{\text{rec}}^{(i)})$. L'ensemble $D_{\text{syn}} = \{D_{\text{syn}}^{(i)} : i = 1, \dots, m\}$ est celui qui est diffusé aux membres du public.

Tout le processus d'imputation appropriée extraite de l'abondante littérature sur l'imputation multiple peut être utilisée pour générer D_{syn} à partir de D_{rec} . Les méthodes pour population finie proposées ici peuvent être appliquées que l'on ait ou non supposé qu'une population finie a été utilisée pour générer D_{syn} . Sous une hypothèse de population finie, puis que les données sont entièrement observées (recensement), les paramètres du modèle d'imputation seraient considérés comme étant connus et fixes. Voir Reiter et Kinney (2012) pour une illustration de la façon d'obtenir des inférences valides à partir d'échantillons aléatoires

problèmes de très grande portée et très complexes, dans certains cas nettement plus complexes que ceux qui se posent habituellement dans le secteur de la statistique officielle.

Remerciements

Nous remercions M. Philip Kokic de la Commonwealth Scientific and Industrial Research Organisation, de nous avoir fourni les données sur les exploitations agricoles à grande échelle (*broodacre farms*). Nous remercions aussi un rédacteur associé et les examinateurs de leurs commentaires constructifs au sujet de la version originale du présent article.

Bibliographie

ABARE (2003). Australian farm surveys report 2003. Canberra.

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of American Statistical Association*, 88, 268-277.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2010). Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de distribution finie dans l'échantillonnage avec probabilités intégrales. *Techniques d'enquête*, 36, 1, 25-37.

Craibiceanu, C.M., Ruppert, D., Carroll, R.J., Joshi, A. et Goodner, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic error. *Journal of Computational and Graphical Statistics*, 16, 265-288.

Dorfman, H., et Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1474.

Francisco, C.A., et Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.

Harns, T., et Duchesne, P. (2006). De l'estimation des quantiles par calage. *Techniques d'enquête*, 32, 1, 41-57.

Kuk, A.Y.C. (1993). A kernel method for estimating finite population functions using auxiliary information. *Biometrika*, 80, 385-392.

Kuk, A.Y.C., et Welsh, A.H. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society, Series B*, 63, 277-292.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, DOI: 10.1198/016214504000000467. {70}, 99, 546-556.

Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (avec discussion et réplique). *Journal of Official Statistics*, 28, 309-334.

Lombardía, M.J., González-Manteiga, W. et Prada-Sánchez, J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference*, 116, 367-388.

Lombardía, M.J., González-Manteiga, W. et Prada-Sánchez, J.M. (2004). Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimate of a finite population distribution function. *Journal of Nonparametric Statistics*, 16, 63-90.

Rao, J.N.K., Kovar, J.G. et Manei, H.J. (1990). On estimating distribution function and quantile from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Royal, R.M., et Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Ruppert, D., Wand, M.P. et Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge, Royaume-Uni : Cambridge University Press.

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantiles. *Journal of Official Statistics*, 24, 495-506.

Sitter, R.R., et Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters*, 52, 353-358.

Wang, S., et Dorfman, A.H. (1996). A new estimator for the finite population distribution function. *Biometrika*, 83, 639-652.

Wood, S.N. (1994). Monotonic smoothing splines fitted by cross validation SIAM. *Journal on Scientific Computing*, 15, 1126-1133.

Woodruff, R. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complex auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Zheng, H., et Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zheng, H., et Little, R.J.A. (2005). Inference for the population total from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

calculer, mais n'utilise pas les variables auxiliaires connues dans les unités non échantillonnées. Une autre possibilité consiste à redéfinir l'estimateur CD au moyen du modèle avec splines que nous avons utilisé pour définir l'estimateur PB2SP. Plus précisément, au lieu de supposer que le modèle de régression passe par l'origine, un modèle avec splines est ajusté aux moments d'ordres un et deux de la loi conditionnelle de la variable résultat étudiée sachant la probabilité d'inclusion. L'estimateur CD fondé sur les splines devrait donner des résultats comparables à ceux de l'estimateur PB2SP, et sa variance peut être estimée en utilisant des méthodes de rééchantillonnage.

Dans le contexte de la statistique officielle, les méthodes décrites dans le présent article illustrent les avantages éventuels d'un changement de paradigme pour passer de méthodes fondées sur le plan de sondage à la modélisation bayésienne en vue de produire des inférences ayant de bonnes propriétés fréquentistes. Nos collègues spécialistes de la statistique fondée sur l'échantillonnage probabiliste ont deux grandes objections à ce point de vue.

Premièrement, l'idée d'une approche exagérément fondée sur un modèle – pire encore, bayésienne – des enquêtes probabilistes est mal acceptée, quoique nous mettions ici l'accent sur des méthodes bayésiennes ayant de bonnes propriétés de randomisation. Selon nous, les méthodes probabilistes classiques fondées sur le plan ne fournissent pas l'approche globale nécessaire pour traiter les problèmes complexes qui se posent de plus en plus souvent en statistique officielle. Des choix judicieux de modèles bien calés sont nécessaires pour s'y attaquer. En accordant de l'attention aux caractéristiques du plan de sondage et en choisissant des lois a priori objectives, on peut obtenir des inférences bayésiennes exemptes de subjectivité, et comme les hypothèses de modélisation sont explicites, elles peuvent être critiquées et perfectionnées. Voir Little (2004, 2012) pour une discussion plus approfondie de ces points.

La deuxième objection est que les méthodes bayésiennes requièrent des calculs trop compliqués pour le secteur de la statistique officielle qui doit calculer correctement et produire rapidement un grand nombre de statistiques régulières. Il est vrai qu'à l'heure actuelle, le calcul bayésien peut sembler rébarbatif aux statisticiens habitués à de simples statistiques pondérées et à des méthodes d'estimation de la variance par rééchantillonnage. Dans un article défendant vigoureusement les approches bayésiennes, Sedransk (2008) mentionne que les difficultés pratiques de calcul sont un inhibiteur. Nous convenons que du travail reste à faire pour répondre à cette objection, mais nous ne pensons pas que le problème soit insurmontable. La recherche sur les méthodes de calcul bayésien a connu une véritable explosion ces dernières décennies, tout comme la capacité de calcul. Des modèles bayésiens ont été ajustés pour résoudre des

Pour les données dont la distribution est normale, nous recommandons d'utiliser l'estimateur PB2SP de préférence aux autres, en raison du biais plus petit, de la REQM plus petite, et de la meilleure couverture et de la plus courte largeur de l'intervalle de confiance. L'estimateur PB2SP et son intervalle de probabilité à posteriori de 95 % sont faciles à obtenir en utilisant l'algorithme proposé par Crainiceanu et coll. (2007), qui offre aussi l'avantage d'un temps de calcul relativement court.

L'estimateur PB2SP peut être biaisé quand l'hypothèse de normalité conditionnelle devient plus raisonnable. L'estimateur PB2SP peut être appliqué aux données transformées et les tirages à partir des lois à posteriori des unités non échantillonnées sont de nouveau transformés pour revenir à l'échelle originale avant d'estimer les quantiles d'intérêt.

Dans nos simulations avec des données non normales, l'approche bayésienne avec fonction de répartition inverse demeurait plus efficace que l'estimateur pondéré par les poids de sondage. L'amélioration de la couverture de l'intervalle de confiance était limitée aux situations où la taille de l'échantillon est petite, avec une méthode de détermination de l'IC de Woodruff dominant de bons résultats quand l'hypothèse de grand échantillon est vérifiée. Donc, pour les données ne suivant pas une loi normale pour lesquelles il n'existe aucune transformation évidente en vue d'améliorer la normalité, nous ne recommandons pas l'approche bayésienne avec fonction de répartition inverse quand l'échantillon est de grande taille. Étant donné les bonnes propriétés de l'estimateur PB2SP dans les conditions de normalité, l'extension à examiner lors de futurs travaux consisterait à relâcher l'hypothèse de normalité dans nos approches pondosées.

Nous utilisons la probabilité d'inclusion comme variable auxiliaire ici. Lorsqu'il n'existe qu'une seule variable auxiliaire pertinente, peu importe que l'on modélise la probabilité d'inclusion ou la variable auxiliaire. Par contre, s'il existe plus d'une variable auxiliaire pertinente, la probabilité d'inclusion est la variable auxiliaire principale qui doit être modélisée correctement, puisque la spécification incorrecte du modèle reliant le résultat étudié à la probabilité d'inclusion entraîne un biais. Lorsque d'autres variables auxiliaires sont observées pour toutes les unités de la population finie, nos estimateurs bayésiens peuvent tous deux être étendus facilement afin d'inclure les covariables auxiliaires supplémentaires en ajoutant des termes linéaires pour ces variables dans le modèle avec splines pénalisées correspondant.

Un examinateur a proposé une approche pondérée de rechange fondée sur la loi de Dirichlet, qui est facile à

Tableau 4
Biais empirique $\times 10^2$, racine carrée de l'erreur quadratique moyenne $\times 10^2$, largeur moyenne de l'IC à 95 % $\times 10^2$, et taux de non-ouverture de l'IC à 95 % $\times 10^3$ de $\theta(\alpha)$ pour $\alpha = 0,1, 0,25, 0,5, 0,75$ et $0,9$: données sur les exploitations agricoles à grande échelle

	0,1	0,25	0,5	0,75	0,9
<i>Biais empirique</i>					
FR inverse	8	14	-63	-22	-60
PB2SP	-110	-125	-12	-88	
PS	20	-19	-17	-21	-61
<i>REQM empirique</i>					
FR inverse	117	117	108	164	256
PB2SP	113	141	124	140	206
PS	132	173	167	226	350
<i>Largeur moyenne de l'IC à 95 %</i>					
FR inverse	402	443	501	697	906
PB2SP	170	327	539	726	964
PS	285	468	615	864	1 589
<i>Taux de non-ouverture de l'IC à 95 %</i>					
FR inverse	96	53	26	52	90
PB2SP	670	258	42	8	17
PS	220	121	68	42	44

Les simulations donnent à penser que les deux estimateurs fondés sur un modèle bayésien avec splines donnent de meilleurs résultats que l'estimateur pondéré par les poids de sondage, les estimateurs par le ratio et par la différence fondés sur le plan de sondage, ainsi que l'estimateur CD fondé sur un modèle lorsque le modèle supposé est incorrect. Les nouvelles méthodes donnent toutes deux des racines de l'erreur quadratique moyenne plus petites qu'il n'y ait pas d'association ou qu'il y ait une association linéaire ou une association non linéaire entre le résultat de l'enquête et la probabilité d'inclusion. Dans certains scénarios, l'accroissement de l'efficacité obtenu en utilisant les deux méthodes bayésiennes est considérable. Lorsque l'hypothèse de normalité du résultat étudié sachant les probabilités d'inclusion est vérifiée, l'estimateur PB2SP produit une REQM plus petite et un intervalle de crédibilité plus court que l'approche avec fonction de répartition inverse. En outre, les deux estimateurs fondés sur un modèle bayésien sont robustes à l'erreur de spécification tant de la fonction moyenne que de la fonction variance. En revanche, l'estimateur fondé sur un modèle CD est biaisé et inefficace quand la fonction moyenne ou la fonction variance est mal spécifiée. Enfin, les méthodes fondées sur un modèle bayésien ont l'avantage de permettre de calculer plus facilement l'IC à 95 % et l'intervalle fondé sur les lois a posteriori des paramètres. Cette caractéristique est intéressante, parce que l'estimation de la variance pour les autres estimateurs fondés sur le plan de sondage peut être compliquée. La méthode d'estimation de la variance de Woodruff pour l'estimateur pondéré par les poids de sondage donne de bons résultats quand une fraction

importante des données est sélectionnée à partir de la population finie, même dans les parties moyennes à extrême des queues de la fonction de répartition. Cependant, lorsque les données provenant de la population sont peu nombreuses, la méthode de Woodruff a tendance à sous-estimer la couverture de l'intervalle de confiance, alors que les deux méthodes bayésiennes donnent une couverture de ces intervalles plus proche du niveau nominal. Les trois estimateurs fondés sur le plan de sondage ont un biais empirique global comparable à celui des deux estimateurs fondés sur un modèle bayésien avec splines. Toutefois, la variation du biais de l'estimateur pondéré par les poids de sondage présente une tendance linéaire lorsqu'on augmente la moyenne d'échantillon des probabilités d'inclusion. En l'absence d'association entre le résultat étudié et la probabilité d'inclusion, les estimateurs par le ratio et par différence donnent un biais et une REQM relativement plus grands que l'estimateur pondéré par les poids de sondage. Cependant, dans certains scénarios de simulation, les estimateurs par le ratio et par différence produisent une REQM plus petite que l'estimateur pondéré par les poids de sondage. La comparaison entre l'estimateur pondéré par les poids de sondage classique et l'estimateur pondéré par les poids de sondage lisse cubique lisse pour la fonction de répartition pondérée par les poids de sondage peut améliorer l'efficacité, mais que l'estimateur pondéré par les poids de sondage lisse continuera d'avoir une REQM plus grande que l'estimateur bayésien avec fonction de répartition inverse.

praticiens des sondages. Bien qu'ils soient faciles à calculer et puissent fournir des inférences valides en grand échantillon, les estimateurs pondérés avec un intervalle de confiance de Woodruff peuvent être inefficaces et donner une mauvaise couverture des intervalles de confiance pour les échantillons de taille petite à modérée. Les estimateurs fondés sur un modèle peuvent améliorer l'efficacité des estimations quand le modèle est spécifié correctement, mais produisent des estimations biaisées s'il est mal spécifié. Pour trouver un compromis entre la robustesse et l'efficacité, nous avons considéré des estimateurs fondés sur des modèles avec splines. Pour l'estimation des quantités d'une variable étudiée continue, nous pouvons estimer des fonctions de répartition fondées sur le modèle puis inverser ces fonctions pour obtenir les quantités, ou modéliser directement la variable étudiée sur les probabilités d'inclusion. Dans le présent article, nous proposons deux estimations des quantités fondées sur un modèle bayésien avec splines. La première méthode est celle de l'estimateur bayésien avec fonction de répartition (FR) inverse, obtenue en inversant les estimations fondées sur un modèle avec splines des fonctions de répartition. La deuxième méthode est celle de l'estimateur PBRSP, estimée en supposant que la variable résultat étudiée continue suit une loi normale dont la fonction moyenne et la fonction variance sont toutes deux modélisées au moyen de splines.

L'usage des estimateurs des quantiles de population finie pondérés par les poids de sondage est très répandu chez les

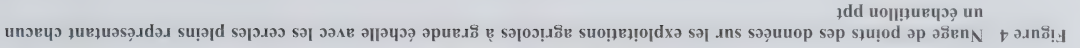


Figure 4 Nuage de points des données sur les exploitations agricoles à grande échelle avec les cercles pleins représentant chacun un échantillon ppt

Tableau 3 Comparaisons de la largeur moyenne et du taux de non- couverture de l'IC à 95 % de $\theta(\alpha)$ pour $\alpha = 0,1, 0,25, 0,5, 0,75$ et $0,9$

Largeur moyenne de l'IC à 95 %	Taux de non- couverture de l'IC à 95 %				
	0,1	0,25	0,5	0,75	0,9

<i>Erreurs homoscedastiques</i>					
NULL	199	156	141	152	184
FR inverse	178	134	118	134	177
PS	195	164	151	167	237
LINUP	257	207	157	139	141
FR inverse	230	167	134	123	121
PS	248	231	188	179	187

<i>Erreurs heteroscedastiques</i>					
NULL	234	184	163	177	234
FR inverse	217	157	132	144	156
PS	231	199	175	210	402
LINUP	257	207	157	139	141
FR inverse	230	167	134	123	121
PS	248	231	188	179	187

NULL	146	104	90	101	137	42	43	38	38	47
FR inverse	107	89	79	89	107	42	49	37	68	65
PS	146	101	91	113	169	80	60	51	37	42
LINUP	131	107	104	124	154	70	31	36	42	40
FR inverse	125	97	87	93	116	47	35	50	58	52
PS	141	110	133	184	219	138	69	41	50	42
EXP	131	99	99	134	242	63	49	34	40	41
FR inverse	116	92	84	98	139	57	55	40	63	59
PS	135	100	106	186	378	111	65	46	45	34

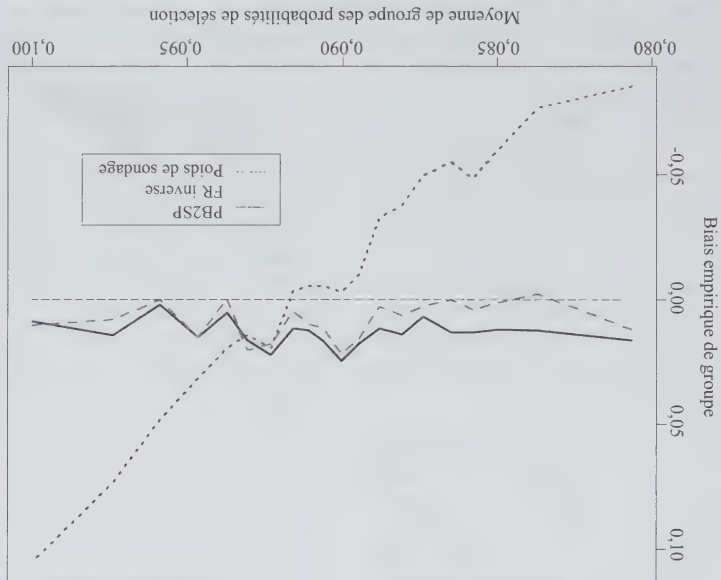
Figure 3 Variation du biais empirique des trois estimateurs pour le 90^e centile dans le cas « EXP + homoscedasticité »

Tableau 1

Comparaisons du biais empirique et de la racine carrée de l'erreur quadratique moyenne $\times 10^3$ de $\theta(\alpha)$ pour $\alpha = 0,1, 0,25, 0,5, 0,75$ et $0,9$: scénarios avec erreurs homoscédastiques

Biais empirique									
REQM empirique									
0,1	0,25	0,5	0,75	0,9	0,1	0,25	0,5	0,75	0,9

NULL	FR inverse	-6	-3	-1	-1	46	37	36	37
	PB2SP	-5	-1	1	2	41	33	31	42
	PS	-5	-3	-1	-4	54	39	39	50
	PS lisse	-7	-4	-1	-2	50	39	37	47
	CD	-197	-272	-265	-108	203	274	266	189
LINUP	Ratio de RKM	3	25	33	16	6	77	125	112
	Dif de RKM	-5	-1	6	14	58	58	94	122
	FR inverse	-15	-3	-2	-1	70	49	49	33
	PB2SP	-3	-1	4	7	56	43	35	31
	PS	-15	-3	-2	-6	77	57	48	44
EXP	Ratio de RKM	-14	-5	-2	-1	72	53	45	42
	Dif de RKM	-101	-35	-37	-49	104	38	39	53
	FR inverse	-23	-9	2	5	95	67	53	51
	PB2SP	-9	-4	-2	-7	65	49	47	47
	PS lisse	-12	-5	-2	-2	62	47	43	46
EXP	Ratio de RKM	-17	-11	1	3	87	65	65	55
	Dif de RKM	-9	-4	-2	-2	60	45	41	43
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
EXP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4	4	7	43	49	41	49
	PB2SP	-10	-6	-3	0,3	52	40	35	36
	PS	-9	-3	-2	-8	65	49	46	50
LINUP	Ratio de RKM	-10	-6	-3	0,3	52	40	35	36
	Dif de RKM	-10	-6	-3	0,3	52	40	35	36
	FR inverse	-8	0,4						

nous attendre à ce qu'elle donne de meilleurs résultats que l'approche PB2SP lorsque l'hypothèse de normalité est violée. Cela motive une comparaison de l'estimateur pondéré par les poids de sondage et de l'estimateur avec fonction de répartition inverse pour des données ne suivant pas une loi normale.

La population considérée ici est définie par 398 exploitations agricoles à grande échelle (qui produisent des céréales, des bovins, des moutons et de la laine) ayant une superficie agricole de 6 000 hectares ou moins qui ont participé à l'Australian Agricultural and Grazing Industries Survey de 1982 réalisée par l'Australian Bureau of Agricultural and Resource Economics (ABARE 2003). La variable Y est le total des recettes monétaires agricoles.

Nous avons tiré 1 000 échantillons systématiques ppt de

taille égale à 100 en prenant la superficie agricole, X_1 , comme variable de taille, de sorte que les grandes exploitations agricoles sont plus susceptibles que les autres d'être sélectionnées dans l'échantillon. La figure 4 donne le nuage de points de Y en fonction de la variable de taille X_1 pour ces exploitations, chaque cercle plein représentant un échantillon ppt sélectionné. Ce graphique montre que la variation de Y augmente à mesure que X_1 augmente. En outre, la distribution de Y est étalée vers la droite étant donné X_1 . Nous avons réalisé une étude par simulation en utilisant ces données sur les exploitations agricoles à grande échelle pour comparer les deux estimateurs fondés sur un modèle bayésien avec splines à l'estimateur pondéré par les poids de sondage.

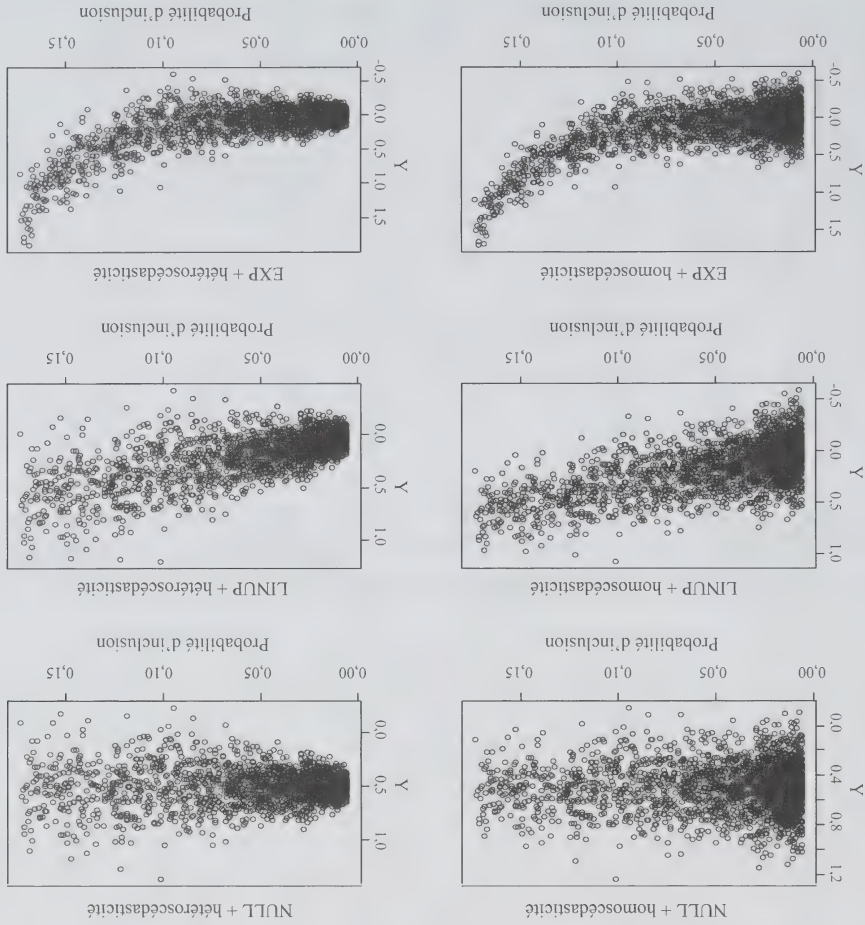


Figure 2 Nuages de points de Y en fonction des probabilités d'inclusion pour les six populations finies artificielles de taille égale à 2 000

variable de taille X_i et $\theta_x(\alpha)$ est le quantile de population connu de X_i ;
 e) Diff, l'estimateur par la différence de RKM (1990) donné par $\hat{\theta}_y(\alpha) + R \times \{\theta_x(\alpha) - \hat{\theta}_x(\alpha)\}$, où R est l'estimation pondérée par les poids de sondage de Y/X .

Les sept estimateurs pour les 10^e, 25^e, 50^e, 75^e et 90^e centiles de la population finie ont été comparés pour ce qui est du biais empirique et de la racine carrée de l'erreur quadratique moyenne (REQM). Étant donné la complexité de l'estimation de la variance des estimateurs CD et RKM, nous avons comparé seulement la largeur moyenne et le taux de non-couverture de l'intervalle de confiance/crédibilité (IC) à 95 % pour les deux estimateurs fondés sur un modèle bayésien et l'estimateur pondéré par les poids de sondage. Pour l'IC à 95 %, nous avons utilisé la méthode de Woodruff pour l'estimateur pondéré par les poids de sondage, la méthode illustrée à la figure 1(c) pour l'estimateur bayésien avec fonction de répartition inverse et la probabilité à posteriori de 95 % du quantile avec queues égales pour l'estimateur PB2SP. Nous avons utilisé des splines cubiques avec 15 nœuds également espacés.

Les tableaux 1 et 2 montrent le biais empirique et la REQM pour les trois distributions normales avec erreurs homoscédastiques et erreurs hétéroscédastiques, respectivement. Dans l'ensemble, le biais empirique dans l'estimation des cinq quantiles est semblable lorsque l'on utilise les estimateurs bayésiens, les deux estimateurs pondérés par les poids de sondage et les deux estimateurs fondés sur le plan de sondage de RKM. Par contre, l'estimateur CD produit un grand biais et une grande REQM dans tous les scénarios, sauf LINUP avec erreur hétéroscédastique, où le modèle sous-jacent de l'estimateur est spécifié correctement. Les deux estimateurs fondés sur un modèle bayésien produisent des racines carrées de l'erreur quadratique moyenne plus petites que les autres estimateurs, et cet accroissement de l'efficacité est important dans certains scénarios, en particulier lorsque l'on utilise l'estimateur PB2SP. Par l'application d'une courbe de régression cubique lisse à la fonction de répartition empirique estimée pondérée par les poids de sondage, l'estimateur pondéré par les poids de sondage lisse produit un gain d'efficacité par rapport aux estimateurs pondérés par les poids de sondage classiques, mais la REQM demeure plus grande que pour l'estimateur bayésien avec fonction de répartition inverse. Les comparaisons des trois estimateurs fondés sur le plan de sondage donnent à penser qu'aucun de ces estimateurs ne domine uniformément les deux autres. En particulier, l'estimateur pondéré par les poids de sondage à une plus petite REQM que les estimateurs par différence et par le ratio de RKM pour les cinq quantiles dans la population NULL et pour les

Bien que l'estimateur pondéré par les poids de sondage se comporte de manière comparable aux estimateurs fondés sur un modèle bayésien avec splines pour ce qui est du biais empirique global, le biais conditionnel des estimations varie considérablement à mesure qu'augmente la moyenne d'échantillon des probabilités d'inclusion. À l'exemple de Royall et Cumberland (1981), nous avons classé les estimations provenant des 1 000 échantillons en fonction de la moyenne d'échantillon des probabilités d'inclusion et nous les avons réparties en 20 groupes de 50, puis nous avons calculé le biais empirique pour chaque groupe. La figure 3 donne le biais conditionnel des deux estimateurs bayésiens et de l'estimateur pondéré par les poids de sondage à mesure qu'augmente la moyenne d'échantillon des probabilités d'inclusion, tandis que le biais de groupe des deux estimateurs fondés sur un modèle bayésien avec splines est moins affecté par cette moyenne. Des comparaisons comparables sont faites pour d'autres scénarios.

3.2 Étude par simulation avec les données de l'enquête sur les exploitations agricoles à grande échelle

L'estimateur PB2SP repose sur l'hypothèse que la variable résultat suit une loi normale, après conditionnement sur les probabilités d'inclusion. Puisque l'approche fondée sur un modèle bayésien avec fonction de répartition inverse ne comporte pas d'hypothèse de normalité, nous pourrions

L'intervalle de crédibilité à 95 % bayésien pour l' α -quantile de population dans les simulations est formé en divisant également la queue de la distribution entre les points finaux supérieur et inférieur.

3. Étude par simulation

3.1 Étude par simulation avec données artificielles

Nous avons d'abord simulé une superpopulation de taille $M = 20\,000$. La variable de taille X dans la superpopulation prend 20 000 valeurs entières consécutives allant de 710 à 20 709. Puis, nous avons tiré une population finie de taille $N = 2\,000$ de cette superpopulation par échantillonnage systématique avec probabilité proportionnelle à la taille (ppt) où la probabilité était proportionnelle à l'inverse de la variable de taille. Par conséquent, dans la population finie, la distribution de la variable de taille est asymétrique avec étallement à droite. La variable résultat étudiée Y a été tirée d'une loi normale de moyenne $f(\pi)$ et de variance d'erreur égale à 0,04 (erreur homoscédastique) ou π (erreur hétéroscédastique). Trois structures de moyenne $f(\pi)$ ont été simulées : pas d'association entre Y et π (NUL), $f(\pi) = 0,5$, une association linéaire (LINUP), $f(\pi) = 6\pi$, et une association non linéaire (EXXP), $f(\pi) = \exp(-4,64 + 52\pi)$. Pour chacune des six conditions de simulation, nous avons généré un millier de répétitions de la population finie et nous avons tiré de chaque population un échantillon ppt systématique ($n = 100$) avec x comme variable de taille ; donc $\pi_i = nx_i / \sum_{j=1}^n x_j$. Les nuages de points de Y en fonction de π pour ces six populations sont présentés à la figure 2.

Nous avons comparé les propriétés de l'estimateur bayésien PB2SP à cinq autres approches :

- a) PS, l'estimateur pondéré par les poids de sondage défini par inversion de F_w^* ;
- b) PS lisse, l'estimateur pondéré par les poids de sondage lisse. Une courbe de régression cubique lisse a été ajustée à F_w^* et désignée par F_w .
- c) CD, l'estimateur de Chambers et Dunstan (1986), en supposant le modèle suivant : $Y_i = \beta\pi_i + U_i$, où U_i est une variable aléatoire dont les valeurs sont indépendantes et identiquement distribuées de moyenne nulle ;
- d) Ratio, l'estimateur par le ratio de RKM (1990) donné par $\{\theta_j(\alpha) / \theta^x(\alpha)\} \times \theta_j^x(\alpha)$, où $\theta_j^x(\alpha)$ et $\theta^x(\alpha)$ désignent respectivement les estimations pondérées par les poids de sondage pour Y et la

Rupert et coll. (2003) ont proposé une approche itérative pour estimer les paramètres de (6). Ils ont d'abord supposé que SPL_2 était connue et ont ajusté un modèle linéaire mixte pour estimer les paramètres dans SPL_1 . Ils ont calculé le carré de la différence entre Y et SPL_1 , qui suivait une loi gamma de paramètre de forme $\frac{1}{2}$ et de paramètre d'échelle $2SPL_2$. Ils ont ensuite ajusté un modèle linéaire mixte généralisé pour les carrés des différences afin d'estimer les paramètres dans SPL_2 . Ils ont itéré les procédés susmentionnés jusqu'à ce que les estimations des paramètres convergent. Cette approche itérative est simple à mettre en œuvre. Cependant, ici, notre objectif n'est pas d'estimer les paramètres, mais d'obtenir des prédictions bayésiennes de Y pour les unités non échantillonnées afin de pouvoir utiliser (4) pour estimer les quantiles.

Crainiceanu, Rupert, Carroll, Joshi et Goodner (2007) ont élaboré une méthodologie inférencielle bayésienne pour (6). Ils ont constaté que la mise en œuvre de la méthode MCMC en utilisant des pas de Metropolis-Hastings multi-variés est instable avec de mauvaises propriétés de mélange. Ils ont proposés d'ajouter des termes d'erreur à la deuxième spline pour rendre les calculs plus faisables, en remplaçant l'échantillonnage à partir de lois conditionnelles complètes complexes par de simples pas de Metropolis-Hastings univariés. Cette idée peut s'exprimer comme

$$\begin{aligned} Y_i &\sim N(SPL_1(\pi_i, k), \sigma_k^2(\pi_i)), \\ \log(\sigma_k^2(\pi_i)) &\sim N(SPL_2(\pi_i, k), \sigma_k^2). \end{aligned}$$

Nous avons utilisé une loi a priori $N(0, 10^6)$ pour les paramètres à effets fixes β et α , et une loi a priori gamma inverse propre $\text{Gamma}(10^{-6}, 10^6)$ pour les valeurs de la variance τ_k^2 et τ_p^2 . Nous avons fixé les valeurs de $\sigma_k^2 = 0,1$. Les lois conditionnelles à posteriori complètes sont décrites en détail dans Crainiceanu et coll. (2007).

La loi à posteriori de l' α -quantile de population finie est simulée en générant un grand nombre D de tirages et en utilisant l'estimateur prédictif de la forme

$$\hat{\theta}_{(p)}(\alpha) = \inf \left\{ t : N^{-1} \left(\sum_{j=1}^t \Delta(t - y_j) + \sum_{j=1}^J \Delta(t - y_j^{(d)}) \right) \geq \alpha \right\},$$

où $y_j^{(p)}$ est un tirage à partir de la loi prédictive à posteriori de la j^{e} unité non échantillonnée de la variable résultat continue. La moyenne de ces tirages simule l'estimateur prédictif bayésien avec deux moments modélisés par splines pénalisées (PB2SP) de l' α -quantile de population finie.

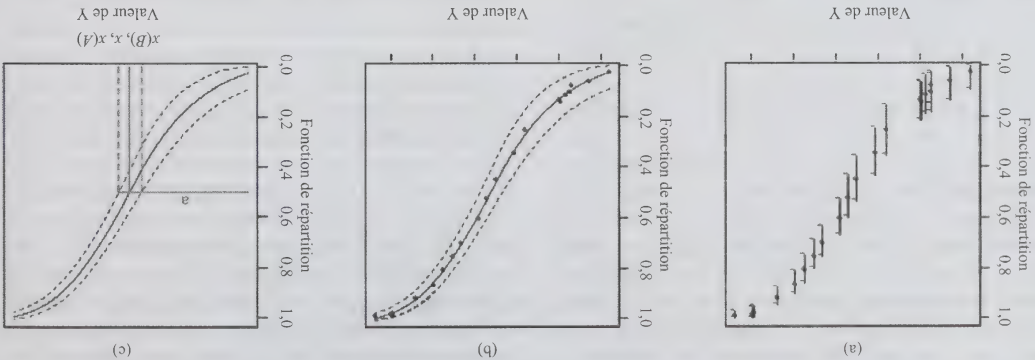


Figure 1 Approche fondée sur un modèle bayésien avec inversion de la fonction de répartition (FR) pour estimer les fonctions de répartition de population finie et les quantiles associés, illustrée en utilisant un échantillon de taille 100 tiré d'une population finie. (a) La méthode PBSP est utilisée pour estimer les fonctions de répartition de la population finie à vingt points de l'échantillon ; les points représentent les estimateurs PBSP et les signes moins représentent les limites supérieure et inférieure des IC à 95 %. (b) Trois modèles de régression cubiques lisses monotones sont ajustés sur les fonctions de répartition continues prédictives et les deux courbes en trait interrompu représentent les IC à 95 % des fonctions de répartition. (c) L'estimation ponctuelle et l'IC à 95 % de l' α -quantile de la population sont obtenus en inversant la fonction de répartition estimée ; x est l'estimation ponctuelle et $x(B)$ et $x(A)$ sont les limites inférieure et supérieure de l'IC à 95 %.

2.2 Approche prédictive bayésienne avec deux moments modélisés par splines pénalisées

Nous considérons d'autres estimateurs des quantiles de population finie de la forme :

$$\hat{\theta}(\alpha) = \inf \left\{ t : N^{-1} \left(\sum_{i \in s} \Delta(t - y_i) + \sum_{i \in s} \Delta(t - \hat{y}_i) \right) \geq \alpha \right\}, \quad (4)$$

où \hat{y}_i est la valeur de la j^{e} unité non échantillonnée prédite par une régression sur les probabilités d'inclusion $\{\pi_i\}$. Un modèle normal de base pour un résultat continu repose sur l'hypothèse d'une fonction moyenne linéaire en $\{\pi_i\}$, c'est-à-dire :

$$Y_i \sim N(\beta_0 + \beta_1 \pi_i, c_i \sigma^2), \quad (5)$$

avec des constantes c_i connues pour modéliser la variance non constante. Ce modèle donne une estimation biaisée de $\theta(\alpha)$ si la relation n'est pas linéaire. Pour estimer les totaux de population finie, Zheng et Little (2003, 2005) ont remplacé dans (5) la fonction moyenne linéaire par une spline pénalisée, et ont supposé que $c_i = \pi_{ik}$ pour une certaine valeur connue de k . Des simulations ont donné à penser que leur estimateur fondé sur un modèle du total de population finie donne de meilleurs résultats que l'estimateur pondéré par les poids de sondage, même quand la structure de variance est mal spécifiée.

Dans (6), la moyenne et le logarithme de la variance sont modélisés par des splines pénalisées (SPL₁) et (SPL₂) sur $\{\pi_i\}$. La modélisation du logarithme de la variance fait en sorte que les estimations de la variance soient positives. Nous permettons des nombres (m_1, m_2) et des emplacements (k, k') différents des nœuds pour les deux splines.

$$\begin{aligned} Y_i &\sim N(\text{SPL}_1(\pi_i, k), \exp(\text{SPL}_2(\pi_i, k'))), \\ \text{SPL}_1(\pi_i, k) &= \beta_0 + \sum_{d=1}^D \beta_d \pi_i^k + \sum_{m=1}^M \gamma_m (\pi_i - k_l)^d, \\ \text{SPL}_2(\pi_i, k') &= \alpha_0 + \sum_{d=1}^D \alpha_d \pi_i^{k'} + \sum_{m=1}^M \nu_m (\pi_i - k_l')^d, \\ \text{ind} \quad b_l &\sim N(0, \tau_b^2), \\ \text{ind} \quad \nu_l &\sim N(0, \tau_\nu^2). \end{aligned} \quad (6)$$

Pour l'estimation des quantiles au lieu du total, il est important de spécifier correctement la structure de la variance afin d'éviter un biais. Par conséquent, nous étendons le modèle avec spline pénalisée de Zheng et Little (2003) en modélisant la moyenne ainsi que la variance en utilisant des splines pénalisées. Le modèle avec deux moments modélisés par splines pénalisées peut s'écrire (Ruppert, Wand et Carroll 2003, page 264) :

la fonction de répartition de $\theta(\alpha)$. Si l'IC à 95 % de la fonction de répartition $F(\cdot)$ est constitué en divisant en parties égales les queues de la distribution a posteriori, l'intervalle formé par x_A et x_B est un IC à 95 % de $\theta(\alpha)$. La preuve en est la suivante : si α est la limite inférieure de l'IC à 95 % de $F(x_A)$, seulement 2,5 % des tirages de $F(x_A)$ dans la distribution a posteriori sont plus petits que α . C'est-à-dire que

$$\Pr(F^{-1}(\alpha) > F^{-1}(F(x_A))) \equiv \Pr(\theta(\alpha) > x_A) = 0,025.$$

De même si α est la limite supérieure de l'IC à 95 % de $F(x_B)$, $\Pr(\theta(\alpha) < x_B) = 0,975$. Par conséquent, il y a une probabilité de 95 % que $\theta(\alpha)$ soit compris entre x_A et x_B dans la distribution a posteriori, étant donné l'échantillon.

Cette approche fondée sur un modèle bayésien avec inversion de la fonction de répartition permet d'éviter de fortes hypothèses de modélisation et peut être appliquée à des distributions normales ou asymétriques. L'estimation de la fonction de répartition à chacune des n unités échantillonnées permet d'utiliser complètement l'information fournie par l'échantillon, mais requiert d'importants calculs ; l'estimation de la fonction de répartition à $k < n$ valeurs réduit le temps de calcul au prix d'une certaine perte d'efficacité. Dans l'approche classique, les quantiles de population sont estimés par inversion de la fonction de répartition empirique non lissée. Nous recommandons d'ajuster une courbe de régression cubique lisse aux fonctions de répartition estimées avant d'inverser la fonction de répartition estimée résultante. Les estimations résultantes des quantiles sont plus efficaces, parce que la courbe lisse exploite l'information provenant de toutes les données. Des simulations dont les résultats ne sont pas présentés ici donnent à penser que la courbe de la fonction de répartition estimée en se basant sur un sous-ensemble bien choisi de k unités échantillonnées est similaire à celle estimée en se basant sur la totalité des unités échantillonnées, mais le temps de calcul est réduit considérablement.

Nous suggérons de choisir le sous-ensemble de k points de données à intervalles égaux dans le milieu de la distribution, et à intervalles plus fréquents dans les extrêmes afin d'améliorer l'estimation de la fonction de répartition dans les queues. Par exemple, dans notre étude par simulation avec un échantillon de taille 100, nous avons estimé les fonctions de répartition à 20 points : les 3 valeurs les plus faibles, les 3 valeurs les plus grandes et 14 autres points uniformément espacés dans le milieu de l'échantillon rangé par ordre de valeur.

Des études par simulation ont indiqué que l'estimateur PBSP est plus efficace que l'estimateur pondéré par les poids de sondage et que l'estimateur par la régression généralisée de la proportion de population finie, avec une couverture des intervalles de confiance plus proches des niveaux nominaux.

Nous employons l'approche PBSP n fois pour estimer $F(t)$ à chaque des valeurs échantillonnées de y , $t = \{y_1, y_2, \dots, y_n\}$. Cet estimateur ne tient pas compte du fait que nous estimons une fonction de répartition complète, et il ne s'agit pas nécessairement d'une fonction monotone. En outre, l'interpolation linéaire des n fonctions de répartition estimées peut mener à une mauvaise estimation de la fonction de répartition de la population finie. Pour contourner ces deux problèmes, nous ajustons une courbe de régression cubique lisse aux n fonctions de répartition estimées en imposant des contraintes de monotonie (Wood 1994). Nous désignons la fonction de répartition estimée résultante par $F(t)$. L'estimateur fondé sur un modèle bayésien de $\theta(\alpha)$, obtenu par inversion de la fonction de répartition (FR), est alors défini comme il suit :

$$\theta_{\text{inv-FR}}(\alpha) = \inf\{t; F(t) \geq \alpha\}. \quad (3)$$

Nous ajustons également deux autres courbes de régression lisse monotone aux limites supérieures et inférieures des intervalles de crédibilité (IC) à 95 % de ces fonctions de répartition estimées, désignées par $F_U(t)$ et $F_L(t)$. Afin de réduire le temps de calcul dans nos études par simulation, nous estimons uniquement la fonction de répartition à $k < n$ points présélectionnés de l'échantillon.

L'idée fondamentale qui sous-tend cette approche est illustrée graphiquement à la figure 1. Supposons qu'un échantillon de taille 100 est tiré d'une population finie. Nous choisissons 20 observations dans l'échantillon et estimons les fonctions de répartition respectives et les IC à 95 % associées en utilisant l'estimateur PBSP. À la figure 1(a), nous représentons les estimations PBSP pour ces 20 observations par des points noirs et les limites inférieure et supérieure de l'IC à 95 %, par des signes « - » que nous relierons par un trait plein. À la figure 1(b), nous ajoutons trois courbes de prédiction lisses monotones en utilisant un trait plein noir pour l'estimation ponctuelle et des traits pointillés noirs pour les limites supérieure et inférieure des IC à 95 %.

À la figure 1(c), nous traçons à travers le graphique une droite horizontale passant par la valeur α sur l'axe des y . Nous lisons x_A et x_B respectivement sur l'axe des x de façon telle que $F_L(x_A) = \alpha$, $F(x) = \alpha$ et $F_U(x_B) = \alpha$. Alors, x est l'estimation bayésienne avec inversion de

son poids de sondage ; on estime l'écart-type du pourcentage d'unités inférieure à l' α -quantile estimé ; on multiplie ensuite l'écart-type estimé par le centile z approprié, puis on l'ajoute et on le soustrait de α pour construire les limites de confiance pour le pourcentage d'unités inférieures à l' α -quantile estimé. Enfin, les valeurs de la variable étudiée correspondant aux limites de confiance du pourcentage d'unités inférieures à l' α -quantile estimé sont lues sur les unités pondérées de la pseudo-population rangées par ordre de taille. L'estimation de la variance du pourcentage d'unités de la pseudo-population dont la valeur est inférieure à l' α -quantile estimé est discutée dans Woodruff (1952). Sitter et Wu (2001) ont montré que les intervalles de Woodruff donnent de bons résultats, mêmes dans les cas modérés à extrêmes des queues de la fonction de répartition. Une autre estimation de la variance a été établie par Francisco et Fuller (1991) en utilisant une version lissée de la version du test de

2.1 Approche fondée sur un modèle bayésien avec inversion de la fonction de répartition estimée

La fonction quantile de population finie est l'inverse de la fonction de répartition (FR) de population finie, définie comme étant $F(t) = N^{-1} \sum_{i=1}^N \Delta(t - y_i)$, où $\Delta(x) = 1$ quand $x \geq 0$ et $\Delta(x) = 0$ ailleurs. Nous pouvons estimer les quantiles de population finie en commençant par construire une estimation prédictive continue et strictement monotone de $F(t)$, en traitant $\Delta(t - y)$ comme une variable de résultat binaire et en appliquant des méthodes d'estimation des proportions en population finie.

En particulier, Chen et coll. (2010) ont proposé un estimateur prédictif bayésien avec splines pénalisées (PBSP) pour les proportions de population finie sous échantillonnage avec probabilités inégales. Ils font la régression de la variable binaire z sur les probabilités d'inclusion dans l'échantillon, en utilisant le modèle de régression probit avec splines pénalisées (2) avec m nœuds fixes pré-sélectionnés :

$$\Phi^{-1}(E(z_j | \beta, b, \pi_j)) = \beta_0 + \sum_p \beta_p \pi_j^p + \sum_m b_m (\pi_j - k_j)^+ \quad (2)$$

Des unités autoreprésentatives sont incluses en prenant $\pi_j = 1$. En supposant que les lois a priori pour β et τ^2 sont non informatives, ils ont simulé des tirages de z pour les unités non échantillonnées à partir des lois prédictives. Un tirage à partir de la loi a posteriori de la proportion de population finie s'obtient alors en calculant la moyenne des unités échantillonnées observées et des tirages d'unités non échantillonnées. Le procédé est répété de nombreuses fois pour simuler la loi a posteriori de la proportion

plus efficaces que l'estimateur pondéré par les poids de sondage quand la variable d'intérêt résultant de l'enquête était approximativement proportionnelle à la variable auxiliaire.

Nous supposons ici que l'on procède à un échantillonnage avec probabilités inégales où les probabilités d'inclusion sont connues pour toutes les unités de la population. Nous élaborons deux estimateurs bayésiens fondés sur un modèle avec splines des quantiles de population finie dans lequel sont intégrées les probabilités d'inclusion. La première méthode consiste à estimer la fonction de répartition d'un certain nombre de valeurs d'échantillon en utilisant des estimateurs prédictifs bayésiens avec splines pénalisées (Chen et coll. 2010). Nous estimons ensuite les quantiles de population finie en prenant l'inverse de la fonction de répartition prédictive. La deuxième méthode consiste à utiliser un estimateur prédictif bayésien avec splines pénalisées à deux moments, qui prédit les valeurs des unités non échantillonnées en se basant sur un modèle normal, dont la moyenne et la variance sont toutes deux modélisées au moyen de splines pénalisées sur les probabilités d'inclusion. Nous comparons la performance de ces deux nouvelles méthodes à celle de l'estimateur pondéré par les poids de sondage, de l'estimateur CD et des estimateurs par le ratio et par la différence de RKM, en réalisant des études par simulation sur des données générées artificiellement et sur des données d'enquêtes agricoles.

2. Estimateurs des quantiles

Soit s un échantillon aléatoire de taille n tiré avec probabilités inégales de la population finie de N unités identifiées selon les probabilités d'inclusion $\{\pi_i, i = 1, \dots, N\}$ que l'on suppose être connues pour toutes les unités avant qu'un échantillon soit tiré. Soit X une variable étudiée continue, pour laquelle les valeurs $\{y_1, y_2, \dots, y_n\}$ sont observées dans l'échantillon aléatoire s . L' α -quantile de X dans la population finie est défini comme étant :

$$\theta(\alpha) = \inf \left\{ t : N^{-1} \sum_{i=1}^N \Delta(t - y_i) \geq \alpha \right\}, \quad (1)$$

où $\Delta(u) = 1$ quand $u \geq 0$ et $\Delta(u) = 0$ autrement. On estime souvent $\theta(\alpha)$ en utilisant l' α -quantile pondéré par les poids de sondage $\hat{\theta}(\alpha) = \inf \{t; F_n^w(t) \geq \alpha\}$, où $F_n^w(t)$ est la fonction de répartition pondérée par les poids de sondage donnée par

$$F_n^w(t) = \left(\sum_{i=1}^n \pi_i^{-1} \Delta(t - y_i) \right) / \left(\sum_{i=1}^n \pi_i^{-1} \right).$$

Woodruff (1952) a proposé une méthode de calcul des limites de confiance pour l' α -quantile pondéré par les poids de sondage. En premier lieu, on obtient une pseudo-population en pondérant chaque unité de l'échantillon par

Inférence bayésienne pour les quantiles de population finie sous échantillonnage avec probabilités inégales

Qixuan Chen, Michael R. Elliott et Roderrick J.A. Little¹

Résumé

Le présent article décrit l'élaboration de deux méthodes bayésiennes d'inférence au sujet des quantiles de variables d'intérêt continues d'une population finie sous échantillonnage avec probabilités inégales. La première de ces méthodes consiste à estimer les fonctions de répartition des variables étudiées continues en ajustant un certain nombre de modèles de régression probit avec splines pénalisées sur les probabilités d'inclusion. Les quantiles de population finie sont alors obtenus par inversion des fonctions de répartition estimées. Cette méthode demande considérablement de calculs. La deuxième méthode consiste à prédire les valeurs pour les unités non échantillonnées en supposant qu'il existe une relation variant de façon lisse entre la variable étudiée continue et la probabilité d'inclusion, en modélisant la fonction moyenne ainsi que de la fonction de variance en se servant de splines. Les deux estimateurs bayésiens fondés sur un modèle avec splines donnent un compromis désirable entre la robustesse et l'efficacité. Des études par simulation montrent que les deux méthodes produisent une racine carrée de l'erreur quadratique moyenne plus faible que l'estimateur pondéré par les poids de sondage et que les estimateurs par le ratio et par différence décrits dans Rao, Kovar et Mantel (RKM 1990), et qu'ils sont plus robustes à la spécification incorrecte du modèle que l'estimateur fondé sur un modèle de régression passant par l'origine décrit dans Chambers et Dunstan (1986). Lorsque la taille de l'échantillon est petite, les intervalles de crédibilité à 95 % des deux nouvelles méthodes ont une couverture plus proche du niveau nominal que l'estimateur pondéré par les poids de sondage.

Mots clés : Analyse bayésienne ; fonction de répartition ; erreurs hétéroscastastiques ; régression avec splines pénalisées ; échantillons.

1. Introduction

Nous considérons l'inférence pour les quantiles d'une variable continue d'une population finie d'après un échantillon sélectionné avec probabilités inégales. Les quantiles de population finie sont habituellement estimés par les quantiles pondérés par les poids de sondage, c'est-à-dire un estimateur de type Horvitz-Thompson. Souvent, dans les sondages, la variable de plan de sondage (ici, la probabilité d'inclusion) ou une variable auxiliaire corrélée est mesurée sur des unités non échantillonnées, et cette information peut être utilisée pour accroître l'efficacité des estimateurs pondérés par les poids de sondage (Zheng et Little 2003 ; Chen, Elliott et Little 2010).

Les méthodes d'utilisation d'information auxiliaire pour estimer les fonctions de répartition en population finie ont fait l'objet d'études approfondies. Chambers et Dunstan (1986) ont proposé une méthode fondée sur un modèle et ont illustré leur approche au moyen d'un modèle de régression linéaire avec ordonnée à l'origine nulle pour une superpopulation. Dans la suite de l'exposé, nous donnons à cet estimateur le nom d'estimateur CD. Dorfman et Hall (1993) ont appliqué l'approche CD en remplaçant le modèle de régression linéaire par un modèle non paramétrique. Lombardía, González-Manteiga et Prada-Sánchez (2003, 2004) ont proposé une approximation par le bootstrap de ces estimateurs fondée sur le rééchantillonnage d'une version

La recherche sur l'utilisation d'information auxiliaire ont proposé des estimateurs par calage. Wu et Sitter (2001), ainsi que Harms et Duchesne (2006) estimateur fondé sur un modèle avec lissages par noyau, et Chambers, Dorfman et Wehrly (1993) ont proposé un variable étudiée sachant la valeur de la variable auxiliaire. méthode du noyau de la distribution conditionnelle de la connue de la variable auxiliaire avec une estimation par la proposé un estimateur à noyau qui combine la distribution pondérée des estimateurs CD et RKM. Kuk (1993) a Wang et Dorfman (1996) ont proposé une moyenne rapport à l'estimateur CD quand le modèle est mal spécifié. le ratio et par différence fondés sur le plan de sondage par (RKM 1990) ont démontré les avantages des estimateurs par fonction de la variable auxiliaire. Rao, Kovar et Mantel distribution conditionnelle des résidus sous forme d'une la question des écarts par rapport au modèle en estimant la (2001) ont également modifié l'approche CD pour résoudre lissée de la distribution empirique des résidus. Kuk et Welsh

1. Qixuan Chen, professeur adjoint, Department of Biostatistics, Columbia University Mailman School of Public Health, 722 West 168 Street, New York, NY 10032. Courriel : qe2138@columbia.edu ; Michael R. Elliott et Roderrick J.A. Little, professeurs, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109. Courriel : mrelli@umich.edu et rlittle@umich.edu.

Liao, D., et Vaillant, R. (2012). Facteurs d'inflation de la variance dans l'analyse des données d'enquêtes complexes. *Techniques d'enquête*, 38, 1, 57-67.

Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.

Marquardt, D.W. (1980). Comment on "A critique on some ridge regression methods" par G. Smith et F. Campbell: "You should standardize the predictor variables in your regression models". *Journal of the American Statistical Association*, 75(369), 87-91.

Scott, A.J., et Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77(380), 848-854.

Silvey, S.D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society*, 31(3), 539-552.

Wood, F.S. (1984). Effect of centering on collinearity and interpretation of the constant. *The American Statistician*, 2, 88-90.

Snec, R.D., et Marquardt, D.W. (1984). Collinearity diagnostics depend on the domain of prediction, and model, and the data. *The American Statistician*, 2, 83-87.

Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.

Theil, H. (1971). *Principles of Econometrics*. New York : John Wiley & Sons, Inc.

Wissmann, M., Tourenburg, H. et Shalabh (2007). Role of categorical variables in multicollinearity in the linear regression model. Rapport technique numéro 008. Department of Statistics, University of Munich. Disponible au http://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf.

Tableau 9
Indices de conditionnement mis à l'échelle les plus grands et proportions de décomposition de la variance qui y sont associées : quand « aucun régime » est la catégorie de référence pour la variable REGIME dans le modèle

Indice de conditionnement mis à l'échelle	Ordonnée à l'origine	Sexe	lip. tot.	régime tout	régime pauvre en calories	régime pauvre en lipides	régime pauvre en glucides	Proportion mise à l'échelle de la variance de		
TYPE1 : MCO	0,005	0,000	0,016	0,949	0,932	0,157	0,200			
TYPE2 : MCP	0,013	0,008	0,020	0,938	0,926	0,189	0,175			
TYPE3 : MCPPS	0,006	0,007	0,013	0,686	0,741	0,027	0,061			
6										

Tableau 10
Indices de conditionnement mis à l'échelle les plus grands et proportions de décomposition de la variance qui y sont associées : quand « tout régime » est la catégorie de référence pour la variable REGIME dans le modèle

Indice de conditionnement mis à l'échelle	Ordonnée à l'origine	Sexe	lip. tot.	régime tout	régime pauvre en calories	régime pauvre en lipides	régime pauvre en glucides	Proportion mise à l'origine de la variance de		
TYPE1 : MCO	0,982	0,001	0,034	0,968	0,831	0,155	0,186			
TYPE2 : MCP	0,982	0,011	0,029	0,968	0,820	0,182	0,160			
TYPE3 : MCPPS	0,897	0,018	-0,006	0,971	0,318	0,014	-0,019			
17										

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs dont les commentaires ont permis d'apporter d'importantes améliorations au texte. Ce travail de recherche a été partiellement financé par la U.S. National Science Foundation (subvention 0617081). Les opinions, découvertes et conclusions ou recommandations exprimées dans ce texte sont celles des auteurs et ne reflètent pas nécessairement celles de la National Science Foundation.

Bibliographie

Beisley, D.A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*, 38(2), 73-77.

Beisley, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York : John Wiley & Sons, Inc.

Beisley, D.A., Kuh, E. et Welch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York : Wiley Interscience.

Cook, R.D. (1984). Comment on demeaning conditioning diagnostics through centering. *The American Statistician*, 2, 78-79.

Elliot, M.R. (2007). Réduction bayésienne des poids pour les modèles de régression linéaire généralisée. *Techniques d'enquête*, 33, 1, 27-40.

Farrar, D.E., et Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.

Fox, J. (1986). *Linear Statistical Models and Related Methods, with Applications to Social Research*. New York : John Wiley & Sons, Inc.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 1, 5-25.

Hendrickx, J. (2010). *perturb: Tools for evaluating collinearity*. R package version 2.04. Adresse URL <http://CRAN.R-project.org/package=perturb>.

Kish, L., et Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society B*, 36(1), 1-37.

Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 3341-3348.

Li, J. (2007b). Regression diagnostics for complex survey data. Thèse de doctorat non-publiée, University of Maryland.

Li, J., et Valliant, R. (2009). Matrice chapeau et effets de levier pondérés par les poids de sondage. *Techniques d'enquête*, 35, 1, 17-27.

Li, J., et Valliant, R. (2011). Detecting groups of influential observations in linear regression using survey data-adapting the forward search method. Festschrift for Ken Brewer. *Pakistan Journal of Statistics*, 27, 507-528.

Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Thèse de doctorat, University of Maryland.

Tableau 8

Résultat de l'analyse de régression : quand « tout régime » est la catégorie de référence pour la variable RÉGIME dans le modèle

Type de régression	ordonnée à l'origine	tot	lhp.	tout régime	régime en calories pauvre	régime en lipides pauvre	régime en glucides
TYPE1	30.25***	3.20***	0.95	-3.03	1.75	2.75	-1.48
MCO	(2.00) ^b	(0.70)	(0.72)	(1.94)	(2.03)	(2.72)	(3.66)
TYPE2	27.52***	3.65***	1.44*	-1.39	4.46*	3.86	0.94
MCP	(1.71)	(0.82)	(0.67)	(1.67)	(1.79)	(2.59)	(4.22)
TYPE3	27.52***	3.65***	1.44*	-1.39	4.46**	3.86	0.94
MCPPS	(1.75)	(0.99)	(0.63)	(1.80)	(1.70)	(3.73)	(3.87)

^a valeur p : *, 0,05 ; **, 0,01 ; ***, 0,005.
^b Les erreurs-types sont indiquées entre parenthèses sous les estimations des paramètres.

Lorsqu'on choisit « aucun régime » comme catégorie de référence pour RÉGIME au tableau 9, les indices de conditionnement mis à l'échelle sont relativement petits et ne signalent aucune quasi-dépendance remarquable, quel que soit le type de régression. Seule la dernière ligne pour l'indice de conditionnement le plus grand est imprimée dans les tableaux 9 et 10. Souvent, la catégorie de référence d'une variable explicative catégorique est choisie de manière qu'elle ait une signification analytique. Dans le présent exemple, l'utilisation de « aucun régime » serait logique.

Au tableau 10, lorsque l'on choisit « tout régime » comme catégorie de référence par la variable RÉGIME, les indices de conditionnement mis à l'échelle augmentent et indiquent un degré modéré de colinéarité (indice de conditionnement supérieur à 10) entre les variables indicatrices des régimes suivis et l'ordonnée à l'origine. En utilisant le tableau des proportions de décomposition de la variance mise à l'échelle, pour les MCP, les variables indicatrices pour « aucun régime » et « régime pauvre en calories » jouent un rôle dans la dépendance dominante avec l'ordonnée à l'origine ; par contre, pour les MCPPS, seule la variable indicatrice pour « aucun régime » joue un rôle dans la quasi-dépendance dominante avec l'ordonnée à l'origine et les trois autres variables de régime suivi sont nettement moins inquiétantes.

5. Conclusion

La dépendance entre les variables explicatives incluses dans un modèle de régression linéaire ajusté sur des données d'enquête affecte les propriétés des estimateurs des paramètres. Les problèmes sont les mêmes que ceux observés pour les données ne provenant pas d'enquêtes : les erreurs-types des estimateurs de pente peuvent être trop grandes et les pentes estimées peuvent avoir un signe illogique. Dans le cas extrême où une colonne de la matrice de plan est exactement une combinaison linéaire d'autres colonnes, les équations d'estimation ne peuvent pas être résolues. Les cas

Les plus intéressants sont ceux où les variables explicatives sont reliées, mais que la dépendance n'est pas exacte. Les diagnostics de colinéarité disponibles dans les routines des logiciels classiques ne conviennent pas entièrement pour les données d'enquête. Tous les diagnostics qui comportent une estimation de la variance doivent être modifiés pour tenir compte des caractéristiques de l'échantillon telles que la stratification, les grappes et la pondération inégale. Le présent article décrit l'adaptation des nombres de conditionnement et des décompositions de variance, qui peuvent être utilisés pour repérer les cas de dépendance non exacte, afin de les appliquer à l'analyse des données d'enquête.

Le nombre de conditionnement d'une matrice de plan pondérée par les poids de sondage $W^{1/2} X X^T W X$ qui doit être inversée lorsque l'on ajuste un modèle linéaire est d'autant plus grande que le nombre de conditionnement est grand. Les valeurs élevées des nombres de conditionnement sont un symptôme de certains des problèmes numériques associés à la colinéarité. Les termes de la décomposition comprennent aussi les « effets de spécification incorrecte » si les erreurs du modèle ne sont pas indépendantes, comme cela serait le cas dans un échantillon de grappes. La variance de l'estimateur d'un paramètre de régression peut aussi s'écrire comme une somme de termes faisant intervenir les valeurs propres de $W^{1/2} X$. Les décompositions de la variance pour différents estimateurs des paramètres doivent être utilisées pour repérer les variables explicatives qui sont corrélées entre elles. Après avoir déterminé quelles variables explicatives sont colinéaires, un analyste peut décider si la colinéarité a des effets suffisamment importants sur un modèle ajusté pour justifier de prendre des mesures. La correction la plus simple consiste à éliminer une ou plusieurs variables explicatives, à réajuster le modèle, et à observer comment les estimations changent. Les outils que nous fournissons ici permettent de le faire d'une manière appropriée pour les modèles de régression pondérés par les poids de sondage.

variables indicatrices les quatre variables indicatrices concernant le régime utilisé dans l'étude précédente que nous désignons à la présente section par « tout régime » (RÉGIME), « régime pauvre en calories » (RÉGIMECAL), « régime pauvre en lipides » (RÉGIMELIP) et « régime pauvre en glucides » (RÉGIMEGLU). Le modèle considéré ici est le suivant :

$$IMC_{hi} = \beta_0 + \beta_{noire} * noire_{hi} + \beta_{LIP.TOT} * LIP.TOT_{hi} + \beta_{RÉGIME} * RÉGIME_{hi} + \beta_{RÉGIMECAL} * RÉGIMECAL_{hi} + \beta_{RÉGIMELIP} * RÉGIMELIP_{hi} + \beta_{RÉGIMEGLU} * RÉGIMEGLU_{hi} + e_{hi} \quad (20)$$

Les tableaux 7 et 8 présentent les résultats de l'analyse de régression du modèle (20) en utilisant les trois types de régression – MCO, MCP et MCPPS – énumérés au tableau 1. Le tableau 7 correspond à la modélisation des effets des facteurs de régime suivi sur l'IMC en traitant « aucun régime » comme la catégorie de référence pour les quatre variables de régime suivi, tandis que le tableau 8 correspond au changement du niveau de référence de la variable RÉGIME pour passer de « aucun régime » à « tout régime » et à la modélisation de l'effet de « aucun régime » sur l'IMC. Le choix du niveau de référence a une incidence sur le signe du coefficient estimé pour la variable RÉGIME, mais non sur sa valeur absolue ni sur son erreur-type. La grandeur de l'ordonnée à l'origine estimée et son erreur-type sont différentes dans les tableaux 7 et 8, mais les fonctions estimables, comme les prédictions, seront naturellement les mêmes pour l'un et l'autre ensemble de niveaux de référence. L'erreur-type de l'ordonnée à l'origine est environ trois fois plus grande lorsque la catégorie « tout régime » est le niveau de référence de la variable RÉGIME (tableau 8) que quand elle ne l'est pas (tableau 7).

Tableau 7
Résultat de l'analyse de régression : quand « aucun régime » est la catégorie de référence pour la variable RÉGIME dans le modèle

type de régression	ordonnée à l'origine	noire	lip.	tot	régime	régime pauvre en calories	régime pauvre en lipides	régime pauvre en glucides
TYPE1	27,22***	3,20***	0,95	3,03	1,75	1,75	2,75	-1,48
MCO	(0,61) ^a	(0,70)	(0,72)	(1,94)	(2,03)	(2,03)	(2,72)	(3,66)
TYPE2	26,13***	3,65***	1,44*	1,39	4,46*	4,46*	3,86	0,94
MCP	(0,58)	(0,82)	(0,67)	(1,67)	(1,79)	(1,79)	(2,59)	(4,22)
TYPE3	26,13***	3,65***	1,44*	1,39	4,46**	4,46**	3,86	0,94
MCPPS	(0,64)	(0,99)	(0,63)	(1,80)	(1,70)	(1,70)	(3,73)	(3,87)

^a valeur p : *, 0,05 ; **, 0,01 ; ***, 0,005.
^b Les erreurs-types sont indiquées entre parenthèses sous les estimations des paramètres.

totaux devient significatif dans le modèle réduit. La réduction du nombre de variables corrélées semble avoir amélioré considérablement l'exactitude de l'estimation de l'effet de l'apport de lipides totaux sur l'IMC. Notons que les diagnostics de colinéarité ne fournissent pas une voie unique vers un modèle final. Le choix des variables explicatives particulières qui doivent être éliminées ou retenues peut varier selon l'analyste.

4.3 Deuxième étude : niveau de référence pour des variables catégoriques

Comme nous l'avons mentionné plus haut, lorsqu'on utilise des données ne provenant pas d'une enquête, les variables indicatrices peuvent aussi jouer un rôle important en tant que source éventuelle de colinéarité. Le choix du niveau de référence pour une variable catégorique peut avoir une incidence sur le degré de colinéarité des données. Plus précisément, choisir comme référence une catégorie dont la fréquence est faible et omettre ce niveau pour ajuster le modèle peut donner lieu à une colinéarité avec le terme d'ordonnée à l'origine. Ce phénomène se transpose à l'analyse des données d'enquête comme nous allons l'illustrer.

lipides est l'une des variables clés que l'analyste estime devoir garder, le sucre pourrait être abandonné pour commencer, suivi par les protéines, les calories, l'alcool, les glucides, les lipides totaux, les fibres alimentaires, les acides gras monoinsaturés totaux, les acides gras polyinsaturés totaux et les acides gras saturés totaux. D'autres remèdes contre la colinéarité pourraient comprendre la transformation des données ou l'utilisation de techniques spécialisées, telles que la régression ridge et la modélisation bayésienne mixte, qui requièrent de l'information (a priori) supplémentaire qui dépasse le cadre de la plupart des travaux de recherche et des évaluations.

Pour démontrer comment les diagnostics de colinéarité peuvent améliorer les résultats de régression dans le présent exemple, le tableau 6 représente les résultats de l'analyse de régression par les MCPPS des modèles originaux contenant toutes les variables explicatives et d'un modèle réduit contenant un moins grand nombre de ces variables. Dans le modèle réduit, toutes les variables d'apport alimentaire sont éliminées sauf l'apport de lipides totaux. Après réduction du nombre de variables corrélées posant problème, l'erreur-type de l'apport de lipides totaux n'est plus que le 46^e de son erreur-type dans le modèle original. L'apport de lipides

Tableau 6
Résultat de l'analyse de régression en utilisant TYPE3 : MCPPS

Variable	Coefficient	E.-T.*	Coefficient	E.-T.*
Ordonnée à l'origine	24,14***	2,77	24,20***	2,69
Âge	0,06	0,08	0,06	0,08
Race noire	3,19***	1,04	3,67***	0,98
Tout régime ^c	1,79	1,52	1,28	1,80
Régime faible en calories	4,09***	1,50	4,59**	1,69
Régime faible en lipides	3,67	2,86	3,87	3,76
Régime faible en glucides	0,46	3,51	0,87	3,86
Calories	-0,88	2,36		
Protéines	7,05	9,59		
Glucides	3,69	9,62		
Sucre	-0,31	1,11		
Fibres alimentaires	-14,52*	5,89		
Alcool	2,09	16,47		
Lipides totaux	29,34	31,37	1,47*	0,68
Acides gras saturés totaux	-15,90	20,18		
Acides gras monoinsaturés totaux	-22,40	23,01		
Acides gras polyinsaturés totaux	-27,69	21,10		
Coefficient p intra-groupe	0,0366		0,0396	

* erreur-type, valeur p : *, 0,05 ; **, 0,01 ; ***, 0,005.

^c La catégorie de référence est « aucun régime » pour toutes les variables de régime étudiées.

Tableau 5 (suite)
Indices de conditionnement et proportions de décomposition de la variance mis à l'échelle : utilisation de TYPE3 : MCPPS

Indice de conditionnement mis à l'échelle	Glucides	Sucre	Fibres alimentaires	Alcool	Lipides totaux	Lipides sat. ^a	Lipides mono. ^a	Lipides poly. ^a
1
2
3
3
4
5
7
8
10
11
13
21
37	.	0,379
165	0,651	0,749	0,615
566	1,008	1,509	0,740	1,036	0,805	0,486	.	0,390

^a Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.

^a Acides gras saturés totaux.
^a Acides gras monoinsaturés totaux.
^a Acides gras polyinsaturés totaux.

Les valeurs des proportions de décomposition sous les MCO et les MCP sont très semblables et aboutissent à la détermination des mêmes variables explicatives comme étant éventuellement colinéaires. Les résultats pour les MCPPS diffèrent quelque peu, comme il est illustré plus bas. Dans le cas des MCO et des MCP, six variables d'apport alimentaire total quotidien – calories, protéines, glucides, alcool, fibres alimentaires et lipides totaux – interviennent dans la quasi-dépendance dominante qui est associée à l'indice de conditionnement mis à l'échelle le plus grand. Quatre variables d'apport quotidien de lipides – lipides totaux, acides gras saturés totaux, acides gras monoinsaturés totaux et acides gras polyinsaturés totaux – interviennent dans la quasi-dépendance secondaire qui est associée au deuxième plus grand indice de conditionnement mis à l'échelle. Les trois tableaux montrent aussi une quasi-dépendance modérée entre l'ordonnée à l'origine et l'âge. L'indice de conditionnement mis à l'échelle associé est égal à 38 pour les MCO, et à 37 pour les MCP et les MCPPS. Cependant, lorsque l'on utilise les MCPPS, le sucre, les acides gras saturés totaux et les acides gras polyinsaturés totaux semblent également intervenir dans la quasi-dépendance dominante comme le montre le tableau 5. Par ailleurs, seulement trois variables d'apport quotidien de lipides – acides gras saturés totaux, acides gras monoinsaturés totaux et acides gras polyinsaturés totaux – interviennent dans la quasi-dépendance secondaire associée au deuxième plus grand indice de conditionnement mis à l'échelle. Donc, lorsqu'on utilise les MCO ou les MCP, l'effet de la quasi-dépendance entre le sucre, les acides gras saturés totaux, les acides gras polyinsaturés totaux et les six variables d'apport nutritionnel total quotidien n'est pas aussi prononcé que dans le cas des MCPPS. Si l'on utilise les

diagnostics conventionnels des MCO ou des MCP pour les MCPPS, on pourrait laisser passer cette quasi-dépendance. Au lieu d'utiliser les indices de conditionnement et la méthode de décomposition de la variance mis à l'échelle (dans les tableaux 3, 4 et 5), un analyste pourrait essayer de déceler les colinéarités en examinant la matrice des coefficients de corrélation non pondérés au tableau 2. Même si la matrice des coefficients de corrélation montre que presque toutes les variables d'apport alimentaire total quotidien sont fortement ou moyennement corrélées par paires, elle ne peut pas être employée pour déceler fiabement les quasi-dépendances entre ces variables quand elles sont utilisées dans une régression. Par exemple, le coefficient de corrélation entre « tout régime » et « un régime pauvre en calories » est assez grand (0,73). Cette quasi-dépendance est associée à un indice de conditionnement mis à l'échelle de 11 (supérieur à 10, mais inférieur au seuil de 30) dans le cas des MCO et des MCP (présentés aux tableaux 3 et 4) et à un indice de conditionnement mis à l'échelle égal à 2 (inférieur à 10) dans le cas des MCPPS (présenté au tableau 5). L'effet de cette quasi-dépendance ne semble pas être très nuisible quelle que soit la méthode de régression utilisée. Par ailleurs, l'alcool est faiblement corrélé à toutes les variables d'apport nutritionnel total quotidien, mais intervient fortement dans la quasi-dépendance dominante présente à la dernière ligne des tableaux 3 à 5. Après avoir diagnostiqué les profits de colinéarité, la correction ordinaire consisterait à éliminer les variables corrélées, à réajuster le modèle et à réexaminer les erreurs types, les mesures de colinéarité et d'autres diagnostics. Il est conseillé d'omettre les X une à la fois en raison des interactions éventuellement complexes entre les variables explicatives. Dans le présent exemple, si l'apport total de

Tableau 4 Indices de conditionnement et proportions de décomposition de la variance mis à l'échelle : utilisation de TYPE2 : MCP

Proportion mise à l'échelle de la variance de									
protéines	calories	régime pauvre en glucides	régime pauvre en lipides	régime pauvre en calories	Alcool	Lipides totaux	Lipides sat. ^b	Lipides mono. ^c	Lipides poly. ^d
1	1	"	"	"	"	"	"	"	"
2	2	"	"	"	"	"	"	0,609	0,992
3	3	"	"	"	"	"	0,347	"	"
3	3	"	"	"	"	"	"	"	"
4	4	"	"	"	"	"	"	"	"
5	5	"	"	"	"	"	"	"	"
7	7	"	"	"	"	"	"	"	"
8	8	"	"	"	"	"	"	"	"
10	10	"	"	"	0,902	0,878	"	"	"
11	11	"	"	"	"	"	"	"	"
13	13	"	"	"	"	"	"	"	"
21	21	"	"	"	"	"	"	"	"
26	26	"	"	"	"	"	"	"	"
37	37	"	"	"	"	"	"	"	"
165	165	"	0,959	0,940	"	"	"	"	0,963
566	566	"	"	"	"	"	"	"	"
Indice de conditionnement mis à l'échelle									

^a Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.

^b Acides gras saturés totaux.

^c Acides gras monoinsaturés totaux.

^d Acides gras polyinsaturés totaux.

Tableau 5 Indices de conditionnement et proportions de décomposition de la variance mis à l'échelle : utilisation de TYPE3 : MCPPS

Proportion mise à l'échelle de la variance de									
protéines	calories	régime pauvre en glucides	régime pauvre en lipides	régime pauvre en calories	tout régime	race noire	âge	Ordonnée à l'origine	Indice de conditionnement mis à l'échelle

1	1	"	"	"	"	"	"	"	"
2	2	"	"	"	"	"	"	"	"
3	3	"	"	"	"	"	"	"	"
3	3	"	"	"	"	"	"	"	"
4	4	"	"	"	"	"	"	"	"
5	5	"	"	"	"	"	"	"	"
7	7	"	"	"	0,461	1,686	"	0,766	"
8	8	"	"	"	"	"	"	"	"
10	10	"	"	"	"	"	"	"	"
11	11	"	"	"	"	"	"	"	"
13	13	"	"	"	"	"	"	"	"
21	21	"	"	"	"	"	"	"	"
26	26	"	"	"	"	"	"	"	"
37	37	"	"	"	"	"	"	"	"
165	165	"	"	"	"	"	"	"	"
566	566	"	"	"	"	"	"	0,318	"
Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.									

^a Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.

^b Acides gras saturés totaux.

^c Acides gras monoinsaturés totaux.

^d Acides gras polyinsaturés totaux.

Tableau 2
Matrice des coefficients de corrélation de la matrice de données X

Âge	Race	Tout	Régime	Régime	Régime	Calories	Protéines	Ciucides	Sucre	Fibres	Alcool	Lipides totaux	Lipides saturés	Lipides mono.	Lipides poly.
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Âge	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Race noire	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tout régime	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Régime pauvre en calories	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Régime pauvre en lipides	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Régime pauvre en glucides	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Calories	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Protéines	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ciucides	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sucre	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Fibres	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Alcool	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lipides totaux	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lipides saturés	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lipides mono.	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Lipides poly.	0,87 ^c	1	1	1	1	1	1	1	1	1	1	1	1	1	1

^a Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.
^b Coefficients de corrélation inférieur à 0,3 sont omises dans ce tableau.
^c Coefficients de corrélation plus grand que 0,3 sont mises en italique dans ce tableau.
^d Acides gras saturés totaux.
^e Acides gras monoinsaturés totaux.
^f Acides gras polyinsaturés totaux.

Tableau 3
Indices de conditionnement et proportions de décomposition de la variance mis à l'échelle : utilisation de TYPE1 : MCO

Indice de conditionnement mis à l'échelle	Ordonnée à l'origine	âge	race noire	tout régime	régime pauvre en calories	régime lipides	régime glucides	calories	protéines
1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1
22	1	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1	1
38	1	1	1	1	1	1	1	1	1
157	1	1	1	1	1	1	1	1	1
581	1	1	1	1	1	1	1	1	1
Indice de conditionnement mis à l'échelle	1	1	1	1	1	1	1	1	1
Glucides	1	1	1	1	1	1	1	1	1
Sucre	1	1	1	1	1	1	1	1	1
Fibres alimentaires	1	1	1	1	1	1	1	1	1
Alcool	1	1	1	1	1	1	1	1	1
Lipides totaux	1	1	1	1	1	1	1	1	1
Lipides sat.	1	1	1	1	1	1	1	1	1
Lipides mono.	1	1	1	1	1	1	1	1	1
Lipides poly.	1	1	1	1	1	1	1	1	1

^a Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.
^b Acides gras saturés totaux.
^c Acides gras monoinsaturés totaux.
^d Acides gras polyinsaturés totaux.

mis à l'échelle sont utilisés pour repérer les variables touchées par une quasi-dépendance. La corrélation intra-grappe des résidus est présentée à la dernière ligne du tableau 6 sous la colonne intitulée « Modèle original ». Dans le modèle utilisé pour les tableaux 3 à 5, $p = 0,0366$ tel qu'il est estimé d'après un modèle avec effets aléatoires pour les grappes. Comme nous l'avons mentionné à la section 3.2, quand p est petit et que l'échantillon est autopondéré, les proportions de décomposition sous MCPPS peuvent être interprétées de la même façon que celles sous MCO. Bien que l'échantillon de la NHANES ne soit pas équilibré, p est petit dans cet exemple et les proportions de décomposition devraient encore fournir des renseignements utiles.

Dans les tableaux 3, 4 et 5, les méthodes de régression pondérées MCP et MCPPS utilisent la matrice de données \mathbf{X} pour obtenir les indices de conditionnement, tandis que la méthode de régression non pondérée, MCO, utilise la matrice de données \mathbf{X} . La valeur la plus grande de l'indice de conditionnement mis à l'échelle pour les méthodes MCP et MCPPS est de 566, c'est-à-dire une valeur un peu plus faible que celle de 581 pour les MCO. Ces deux valeurs sont nettement plus grandes que 30 et indiquent donc une quasi-dépendance importante entre les variables explicatives dans les trois modèles de régression. Des nombres de conditionnement d'une telle grandeur impliquent que l'inverse de la matrice de plan, $\mathbf{X}^T \mathbf{W} \mathbf{X}$, peut être numériquement instable, c'est-à-dire que des faibles variations dans les données x pourraient entraîner des variations importantes dans les éléments de l'inverse.

Les statistiques diagnostiques de ces régressions, y compris les indices de conditionnement et les proportions de décomposition de la variance mises à l'échelle, sont présentées par des points. Soulignons que, dans la décomposition (12), certains termes peuvent être négatifs, de sorte que certaines « proportions » peuvent être supérieures à 1. Cela se produit dans cinq cas au tableau 5. Selon Belsley et coll. (1980), un indice de conditionnement de 10 indique que la colinéarité a un effet modéré sur les erreurs-types ; un indice de 100 indiquerait un effet important. Dans la présente étude, nous considérons qu'une valeur de l'indice de conditionnement mis à l'échelle supérieure à dix est assez grande et qu'une valeur supérieure à 30 est grande et remarquable. En outre, les grandes proportions de décomposition de la variance mises à l'échelle (supérieures à 0,3) associées à chaque grand indice de conditionnement

Tableau 1
Modèles de régression et leurs statistiques de diagnostic de la colinéarité utilisés dans cette première étude expérimentale

Type	Méthode	Matrice var(β_j)	var(β_j)	Matrice pour les indices de conditionnement ^a	Proportion π_h de décomposition de la variance
TYPE1	MCO	$\hat{\sigma}^2(\mathbf{X}^T \mathbf{X})^{-1}$	$\hat{\sigma}^2 \sum_{j=1}^p \frac{n_{2j}^2}{n_{2j}^2} \frac{h_j^2}{n_{2j}^2} / \sum_{j=1}^p \frac{h_j^2}{n_{2j}^2}$	$\mathbf{X}^T \mathbf{X}$	$\frac{n_{2j}^2}{n_{2j}^2} \frac{h_j^2}{n_{2j}^2} / \sum_{j=1}^p \frac{h_j^2}{n_{2j}^2}$
TYPE2	MCP	$\hat{\sigma}^2(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$	$\hat{\sigma}^2 \sum_{j=1}^p \frac{n_{2j}^2}{n_{2j}^2} \frac{h_j^2}{n_{2j}^2} / \sum_{j=1}^p \frac{h_j^2}{n_{2j}^2}$	$\mathbf{X}^T \mathbf{W} \mathbf{X}$	$\frac{n_{2j}^2}{n_{2j}^2} \frac{h_j^2}{n_{2j}^2} / \sum_{j=1}^p \frac{h_j^2}{n_{2j}^2}$
TYPE3	MCPPS	$\hat{\sigma}^2(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{W} \mathbf{R} \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{R} \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1})$	$\hat{\sigma}^2 \sum_{j=1}^p \frac{n_{2j}^2}{n_{2j}^2} \frac{h_j^2}{n_{2j}^2} / \sum_{j=1}^p \frac{h_j^2}{n_{2j}^2}$	$\mathbf{X}^T \mathbf{W} \mathbf{X}$	$\frac{n_{2j}^2}{n_{2j}^2} \frac{h_j^2}{n_{2j}^2} / \sum_{j=1}^p \frac{h_j^2}{n_{2j}^2}$

^a Dans tous les modèles de régression, les paramètres sont estimés par : $\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$.
^b Les valeurs propres de cette matrice sont utilisées pour calculer les indices de conditionnement pour le modèle de régression correspondant.
^c Les termes n_{2j}^2 et h_j^2 proviennent de la décomposition en valeurs singulières de la matrice des données \mathbf{X} .
^d Les termes n_{2j}^2 et h_j^2 proviennent de la décomposition en valeurs singulières de la matrice des données pondérées $\tilde{\mathbf{X}} = \mathbf{W}^{1/2} \mathbf{X}$.
^e Les termes n_{2j}^2 et h_j^2 proviennent de la décomposition en valeurs singulières de la matrice de données pondérées \mathbf{X} . Le terme g_h est l'élément unitaire de la matrice de l'effet de spécification incorrecte \mathbf{G} .

où $\mathbf{z}_h^* = \mathbf{I} / n_h \sum_{i=1}^n \mathbf{z}_i^* e_{hi}^T$ et $\mathbf{z}_h^* = \mathbf{X}_{hi}^T \mathbf{W}_{hi}^m \mathbf{e}_{hi}^T$ avec $\mathbf{e}_{hi}^T = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\boldsymbol{\beta}}_{\text{pps}}$, et la matrice de variance-covariance \mathbf{R} peut être estimée par

$$\mathbf{R} = \sum_{h=1}^H \frac{n_h}{n} \left[\text{Bld}(\text{diag}(\mathbf{e}_{hi}^T \mathbf{e}_{hi}^T)) - \frac{1}{n_h} \mathbf{e}_{hi}^T \mathbf{e}_{hi}^T \right].$$

L'expression (17) est utilisée, entre autres, dans les projections Stata et SUDAAN. L'estimateur $\text{var}_L(\hat{\boldsymbol{\beta}}^{\text{pps}})$ est convergent et approximativement sans biais sous un plan où les grappes sont sélectionnées avec remise (Fuller 2002). L'estimateur donné par (17) est également un estimateur approximativement sans biais sous le modèle de (15) (voir Liao 2010). Puisque l'estimateur $\text{var}_L(\hat{\boldsymbol{\beta}}^{\text{pps}})$ est aussi disponible dans les logiciels, nous l'utiliserons dans les travaux empiriques présentés à la section 4.

En partant de l'expression (12) dérivée à la section 2, on peut écrire la matrice des proportions de décomposition de la variance Π pour $\text{var}_L(\hat{\boldsymbol{\beta}}^{\text{pps}})$ sous la forme

$$\Pi = (\pi^{jk})^{d \times d} = \mathbf{Q}^T \mathbf{Q}^{-1} \quad (19)$$

avec $\mathbf{Q}_L = (\phi^{ij})^{d \times d} = (\mathbf{V} \mathbf{D}^{-2}) \cdot (\mathbf{V}^T \hat{\mathbf{Q}}_L)^T$ et $\hat{\mathbf{Q}}_L$ sur la diagonale principale et des zéros ailleurs.

4. Exemples numériques

À la présente section, nous illustrons les mesures de colinéarité décrites à la section 3 et examinons leur comportement en utilisant des données sur l'apport alimentaire provenant de la National Health and Nutrition Examination Survey (NHANES) de 2007-2008.

4.1 Description des données

Les données sur l'apport alimentaire sont utilisées pour estimer les types et les quantités d'aliments et de boissons consommés durant la période de 24 heures (de minuit à minuit) qui précède l'entrevue, et pour estimer les apports d'énergie, de nutriments et d'autres composantes alimentaires provenant de ces aliments et boissons. La NHANES est réalisée selon un plan d'échantillonnage probabiliste complexe à plusieurs degrés; certains sous-groupes de population sont suréchantillonnés afin d'accroître la fiabilité et la précision des estimations des indicateurs de l'état de santé pour ces groupes. Parmi les personnes qui ont répondu à l'interview sur place au centre d'examen mobile (CEM), environ 94 % ont fourni des renseignements complets sur les apports alimentaires. Les poids de sondage ont été construits en prenant les poids d'échantillon ajustés pour le CEM et en les rajustant en outre pour tenir compte de la

non-réponse supplémentaire et de la différence de répartition selon le jour de la semaine pour la collecte des données sur les apports alimentaires. Ces poids sont plus variables que les poids produits pour le CEM. Le jeu de données utilisé dans notre étude est un sous-ensemble des données de 2007-2008 composé de femmes de 26 à 40 ans ayant répondu à l'enquête. Les observations comportant des valeurs manquantes pour les variables choisies ont été exclues de l'échantillon qui, en bout de ligne, contient 672 réponses complètes. Les poids finaux dans notre échantillon varient de 6 028 à 330 067, avec un ratio de 55 pour 1. Le National Center for Health Statistics des États-Unis recommande que le plan de sélection de l'échantillon s'approche de la sélection stratifiée avec remise de 32 UPB dans 16 strates, avec 2 UPB dans chaque strate.

4.2 Première étude : variables corrélées

Dans une première étude empirique, nous avons considéré un modèle de régression linéaire de l'indice de masse corporelle (IMC) des participants à l'enquête. Les variables explicatives utilisées comprennent deux variables démographiques, l'âge et la race (Noir, non-Noir) de la personne, quatre variables binaires indiquant si la personne suit tout régime spécial, un régime pauvre en calories, un régime pauvre en lipides et un régime pauvre en glucides (la valeur est 1 si la personne suit le régime en question, et 0 autrement), et dix variables d'apport nutritionnel total quotidien, qui sont les quantités totales de calories (100 kcal), de protéines (100 g), de glucides (100 g), de sucre (100 g), de fibres alimentaires (100 g), d'alcool (100 g), de lipides totaux (100 g), d'acides gras saturés (100 g), d'acides gras monoinsaturés (100 g) et d'acides gras polyinsaturés (100 g). Les coefficients de corrélation entre ces variables sont présentés au tableau 2. Notons que les corrélations entre les variables d'apport nutritionnel total quotidien sont souvent grandes. Par exemple, les corrélations de l'apport total de lipides avec les apports totaux d'acides gras saturés, d'acides gras monoinsaturés et d'acides gras polyinsaturés sont de 0,85, 0,97 et 0,93. Trois types de régression ont été spécifiés pour l'échantillon sélectionné afin de faire la démonstration des différents diagnostics. Des renseignements plus détaillés au sujet de ces trois types de régression et de leurs statistiques diagnostiques sont présentés au tableau 1.

TYPE1 : Régression par les moindres carrés ordinaires (MCO) avec estimation de σ^2 ; les statistiques diagnostiques sont obtenues par les méthodes classiques passées en revue à la section 2;

TYPE2 : Régression par les moindres carrés pondérés (MCP) avec estimation de σ^2 et en supposant que $\mathbf{R} = \mathbf{W}^{-1}$; les indices de conditionnement mis à l'échelle sont

En nous inspirant du développement présenté dans Scott et Holt (1982, section 4), nous pouvons réécrire la matrice EFFSL, \mathbf{G}_{sr} , pour un cas particulier de \mathbf{R}_h de façon qu'il soit plus facile de comprendre les proportions de décomposition dans (12). Considérons le cas particulier de (13) avec

$$\text{Cov}^M(\mathbf{e}^h) = \sigma^2 (1 - p) \mathbf{I}_{m_h} + \sigma^2 p \mathbf{1}_{m_h} \mathbf{1}_{m_h}^T$$

où \mathbf{I}_{m_h} est la matrice identité de dimensions $m_h \times m_h$ et $\mathbf{1}_{m_h}$ est un vecteur de m_h valeurs 1. Dans ce cas,

$$\mathbf{X}_T^h \mathbf{W}_h \mathbf{R}_h \mathbf{W}_h \mathbf{X}_T^h = (1 - p) \mathbf{X}_T^h \mathbf{W}_h^2 \mathbf{X}_T^h + \sum_{i \in s_h} m_{hi} \mathbf{X}_T^h \mathbf{W}_h^2 \mathbf{X}_T^h$$

où $\mathbf{X}_{Bhi} = m_{hi}^{-1} \mathbf{1}_{m_{hi}}^T \mathbf{X}_{hi}$. Supposons que l'échantillon est autopoindré, de sorte que $\mathbf{W}_{hi} = \mathbf{W} \mathbf{1}_{m_{hi}}$. Après certaines simplifications, il s'ensuit que

$$\mathbf{G}_{sr} = \mathbf{W} [\mathbf{I}^p + (\mathbf{M} - \mathbf{I})^p] \mathbf{p}$$

où \mathbf{I}^p est la matrice identité de dimensions $p \times p$ et $\mathbf{M} = (\sum_{h=1}^H \sum_{i \in s_h} \mathbf{X}_T^{Bhi} \mathbf{X}_T^{Bhi}) (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}$. Donc, si l'échantillon est autopoindré et que p est très petit, alors $\mathbf{G}_{sr} \approx \mathbf{W} \mathbf{I}^p$ et $\text{Var}^M(\hat{\beta}^{\text{pps}})$ dans (15) sera approximativement la même que la variance sous MCO. S'il en est ainsi, les proportions de décomposition de la variance sous MCPPS seront comparables aux proportions sous MCO. Dans les problèmes de régression, p est souvent petit, puisqu'il s'agit de la

corrélation des erreurs, $\varepsilon_{hi}^h = \mathbf{X}_T^h \beta$, pour différentes unités plutôt que pour les \mathbf{X}_{hi}^h . Cela tient au fait que les effets de plan sont souvent plus faibles pour les coefficients de régression que pour les moyennes – phénomène qui a été constaté pour la première fois par Kish et Frankel (1974). Dans les applications où p est plus grand, les proportions de décomposition de la variance dans (12) demeurent utiles pour déceler la colinéarité, même si les écarts par rapport à l'hypothèse d'indépendance des termes d'erreur du modèle ont un effet sur elles.

Représentons les résidus au niveau de la grappe comme un vecteur, $\mathbf{e}^h = \mathbf{X}_{hi}^h - \mathbf{X}_{hi}^h \hat{\beta}^{\text{pps}}$. L'estimateur de (15) que nous avons considéré découle à partir de considérations fondées sur le plan de sondage. Un estimateur par linéarisation, approprié quand les grappes sont sélectionnées avec remise, est donné par

$$\text{Var}^L(\hat{\beta}^{\text{pps}}) = (\mathbf{X}_T^h \tilde{\mathbf{X}}_T^h)^{-1} \tilde{\mathbf{G}}^L$$

avec l'effet de spécification incorrecte estimé comme étant

$$\tilde{\mathbf{G}}^L = (\tilde{\mathbf{g}}^h)^{d \times d} = \left[\sum_{h=1}^H \frac{m_h}{m} \frac{1}{m_h} \sum_{i \in s_h} (\mathbf{z}_{*i}^{hi} - \mathbf{z}_{*}^{hi})(\mathbf{z}_{*i}^{hi} - \mathbf{z}_{*}^{hi})^T \right] (\mathbf{X}_T^h \tilde{\mathbf{X}}_T^h)^{-1}, \quad (18)$$

d'échantillonnage stratifié à plusieurs degrés, il existe des strates $h = 1, \dots, H$ dans la population, des grappes $i = 1, \dots, N_h$ dans la strate h et des unités $t = 1, \dots, m_{hi}$ dans la grappe hi . Nous sélectionnons les grappes $i = 1, \dots, m_{hi}$ dans la strate h et les unités $t = 1, \dots, m_{hi}$ dans la grappe hi . Désignons l'ensemble de grappes échantillonnées dans la strate h par s_h et l'échantillon d'unités dans la grappe hi par s_{hi} . Le nombre total d'unités échantillonnées dans la strate h est $m_h = \sum_{i \in s_h} m_{hi}$ et le nombre total dans l'échantillon est $m = \sum_{h=1}^H m_h$. Supposons que les grappes sont sélectionnées avec des probabilités variables et avec remise dans les strates et indépendamment entre les strates. Le modèle que nous considérons est :

$$\begin{aligned} E_{sr}(\mathbf{Y}^{hi}) &= \mathbf{x}_T^{hi} \beta \\ h &= 1, \dots, H, \quad i = 1, \dots, N_h, \quad t = 1, \dots, m_{hi} \\ \text{Cov}^M(\varepsilon_{hi}^h, \varepsilon_{h' i'}^h) &= 0 \end{aligned}$$

$$\text{ou} \quad \varepsilon_{hi}^h = \mathbf{X}_{hi}^h \beta, \quad i \neq i'$$

$$\text{Cov}^M(\varepsilon_{hi}^h, \varepsilon_{h' i'}^h) = 0 \quad h \neq h',$$

Nous supposons que les unités sont corrélées dans chaque grappe, mais il n'est pas nécessaire ici de spécifier la forme particulière des covariances pour l'analyse. L'estimateur $\hat{\beta}^{\text{pps}}$ du paramètre de régression peut s'écrire :

$$\hat{\beta}^{\text{pps}} = \sum_H \sum_{i \in s_h} (\tilde{\mathbf{X}}_T^h \tilde{\mathbf{X}}_T^h)^{-1} \mathbf{X}_T^{hi} \mathbf{W}_{hi} \mathbf{Y}^{hi} \quad (14)$$

où \mathbf{X}_{hi}^h est la matrice de dimensions $m_{hi} \times p$ des covariables pour les unités échantillonnées dans la grappe hi , $\mathbf{W}_{hi}^h = \text{diag}(w_i^h)$, $i \in s_{hi}$, est la matrice diagonale des poids de sondage pour les unités dans la grappe hi et \mathbf{Y}^{hi} est le vecteur de dimension $m_{hi} \times 1$ des variables réponses dans la grappe hi . La variance sous le modèle de $\hat{\beta}^{\text{pps}}$ est :

$$\text{Var}^M(\hat{\beta}^{\text{pps}}) = (\tilde{\mathbf{X}}_T^h \tilde{\mathbf{X}}_T^h)^{-1} \mathbf{G}_{sr} \quad (15)$$

où

$$\mathbf{G}_{sr} = \sum_H \sum_{i \in s_h} \sum_{i' \in s_h} \mathbf{X}_T^{hi} \mathbf{W}_{hi} \mathbf{R}_{hi} \mathbf{W}_{h'i'} \mathbf{X}_T^{h'i'} = \sum_H \sum_{i \in s_h} \mathbf{X}_T^{hi} \mathbf{W}_{hi} \mathbf{R}_{hi} \mathbf{W}_{hi} \mathbf{X}_T^{hi} \quad (16)$$

avec $\mathbf{R}_{hi} = \text{Var}_{sr}(\mathbf{Y}^{hi})$, $\mathbf{W}_{hi}^h = \text{diag}(\mathbf{W}_{hi})$, et $\mathbf{R}_{hi} = \text{Bldiag}(\mathbf{R}_{hi})$, $\mathbf{W}_{hi}^h = \text{diag}(\mathbf{W}_{hi})$, $\mathbf{X}_T^{hi} = (\mathbf{X}_T^{h1}, \mathbf{X}_T^{h2}, \dots, \mathbf{X}_T^{h, m_{hi}})$, $\mathbf{X}_T^{hi} = (\mathbf{X}_T^{h1}, \mathbf{X}_T^{h2}, \dots, \mathbf{X}_T^{h, m_{hi}})$, où \mathbf{X}_{hi}^h est la matrice de dimensions $m_{hi} \times p$ des covariables pour les unités échantillonnées dans la strate h , $\mathbf{W} = \text{diag}(\mathbf{W}_{hi})$, pour $h = 1, \dots, H$ et $i \in s_h$ et $\mathbf{R} = \text{Bldiag}(\mathbf{R}_{hi})$.

La variance sous le modèle de l'estimateur MCPPS des paramètres sous un modèle avec $\text{Var}_M(\epsilon) = \sigma^2 \mathbf{R}$ est donnée par :

$$\text{Var}_M(\beta^{\text{PPS}}) = \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{R} \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{G}, \quad (8)$$

où

$$\mathbf{G} = (\mathbf{g}_{ij})^{p \times p} = \mathbf{X}^T \mathbf{W} \mathbf{R} \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \quad (9)$$

est la matrice d'effet de spécification incorrecte (EFFSI) qui représente le facteur d'inflation nécessaire pour corriger les résultats standard afin de tenir compte de l'effet de la corrélation intra-grappe dans les données d'enquête en grappes et du fait que $\text{Var}_M(\epsilon) = \sigma^2 \mathbf{R}$ et non $\sigma^2 \mathbf{W}^{-1}$ (Scott et Holt 1982).

En utilisant la décomposition en valeurs singulières de \mathbf{X} , nous pouvons réécrire $\text{Var}_M(\beta^{\text{PPS}})$ sous la forme

$$\text{Var}_M(\beta^{\text{PPS}}) = \sigma^2 \mathbf{V} \mathbf{D}^{-2} \mathbf{V}^T \mathbf{G}. \quad (10)$$

Le k^{e} élément diagonal dans $\text{Var}_M(\beta^{\text{PPS}})$ est la variance estimée du k^{e} coefficient, β_k . En utilisant (10), on peut exprimer $\text{Var}_M(\beta_k)$ comme :

$$\text{Var}_M(\beta_k) = \sigma^2 \sum_{j=1}^J \frac{\pi_j^2}{V_{kj}^2} \lambda_{kj}^{1/2} \quad (11)$$

où $\lambda_{kj} = \sum_{p=1}^P V_{pj}^2 g_{kp}^2$. Si $\mathbf{R} = \mathbf{W}^{-1}$, alors $\mathbf{G} = \mathbf{I}$, $\lambda_{kj} = V_{kj}$ et (11) se réduit à (3). Cependant, la situation est plus compliquée quand \mathbf{G} n'est pas la matrice identité, c'est-à-dire quand le plan de sondage complexe affecte la variance d'un coefficient de régression estimé. Si les variables explicatives k et j sont orthogonales, $V_{kj} = 0$ pour $k \neq j$ et, dans (11), la variance dépend uniquement de la k^{e} valeur singulière et n'est pas affectée par les g_{ij} qui ne sont pas nuls. Si la variable explicative k et plusieurs variables explicatives j ne sont pas orthogonales, λ_{kj} reçoit la contribution de tous ces vecteurs propres, et des éléments hors diagonale de la matrice EFFSI, \mathbf{G} . Le terme λ_{kj} mesure alors à la fois la non-orthogonalité des x et les effets du plan de sondage complexe.

Par conséquent, nous pouvons définir des proportions de décomposition de la variance analogues à celles obtenues pour les MCO, mais leur interprétation est moins facile. Soit $\phi_{kj} = \lambda_{kj} / \pi_j^2$, $\phi_k = \sum_{j=1}^J \phi_{kj}$ et $\mathbf{Q} = (\phi_{kj})^{p \times p} = (\mathbf{V} \mathbf{D}^{-2}) (\mathbf{V}^T \mathbf{G})^T$. Les proportions de décomposition de la variance sont π_j^2 / ϕ_j , qui représentent la proportion de la variance du k^{e} coefficient de régression associé à la j^{e} composante de sa décomposition dans (11). Désignons la matrice des proportions de décomposition de la variance par

$$\Pi = (\pi_j^2)^{p \times p} = \mathbf{Q}^T \mathbf{Q}^{-1}, \quad (12)$$

où \mathbf{Q} est la matrice diagonale qui contient les sommes de ligne de \mathbf{Q} sur la diagonale principale et des 0 ailleurs. L'interprétation des proportions dans (12) n'est pas aussi catégorique que pour les MCO en raison de l'effet de la matrice EFFSI. À la section 3.2, nous discutons plus en détail de l'interprétation dans le contexte de l'échantillonnage en grappes stratifié.

Comme pour la méthode de régression par les MCO, on peut créer un tableau de décomposition de la variance semblable à celui qui figure à la fin de la section 2. Si deux variables indépendantes ou plus sont collinéaires (ou « quasi dépendantes »), une valeur singulière devrait faire une contribution importante à la variance des estimations du paramètre associées à ces variables. Par exemple, le fait que les proportions π_{j1} et π_{j2} pour les variances de β^{PPS1} et de β^{PPS2} sont grandes indique que la contribution de la troisième valeur singulière aux deux variances est importante et que les première et deuxième variables indépendantes de la régression sont, dans une certaine mesure, collinéaires. Comme il est montré à la section 2.3, quand les k^{e} et j^{e} colonnes de \mathbf{X} sont orthogonales, $V_{kj} = 0$ et la proportion de décomposition de la j^{e} valeur singulière π_j sur $\text{Var}(\beta_k)$ sera égale à 0.

Plusieurs cas particuliers méritent d'être mentionnés. Si $\mathbf{R} = \mathbf{W}^{-1}$ comme il est supposé dans les moindres carrés pondérés (MCP), alors $\mathbf{G} = \mathbf{I}$. La décomposition de la variance en (11) est de la même forme que l'expression (2) pour les MCO. Cependant, il serait inhabituel que $\mathbf{R} = \mathbf{W}^{-1}$ dans des données d'enquête puisque les poids de sondage ne sont généralement pas calculés en se fondant sur la structure de variance d'un modèle. Notons que \mathbf{V} reste différente de celle pour les MCO et est une composante de la décomposition en valeurs singulières de \mathbf{X} au lieu de \mathbf{X} . Un autre exemple est celui où $\mathbf{R} = \mathbf{I}$ et où les poids de sondage sont égaux, auquel cas on peut utiliser les résultats des MCO. Cependant, si les poids de sondage sont inégaux, même quand $\mathbf{R} = \mathbf{I}$, la décomposition de la variance donnée par (11) est différente de celle donnée par (2) dans les MCO, puisque $\mathbf{G} \neq \mathbf{I}$. À la section suivante, nous examinons certains modèles spécifiques qui tiennent compte des caractéristiques de la population, telles que les grappes et les strates, pour estimer cette décomposition de la variance.

3.2 Décomposition de la variance pour un modèle avec mise en grappes stratifiée

La variance sous le modèle de β^{PPS} dans (8) contient la matrice \mathbf{R} inconnue qu'il faut estimer. À la présente section, nous présentons un estimateur de β^{PPS} approprié pour un modèle avec mise en grappes stratifiée. L'estimateur de variance possède une justification sous le modèle ainsi que sous le plan. Supposons que, dans un plan

3. Adaptation aux moindres carrés pondérés par les poids de sondage

3.1 Indices de conditionnement et proportions de décomposition de la variance

Dans le cas des moindres carrés pondérés par les poids de sondage (MCPPS), nous sommes davantage intéressés par les relations de colinéarité entre les colonnes de la matrice $\mathbf{X} = \mathbf{W}^{1/2} \mathbf{X}_0$ qu'entre celles de \mathbf{X}_0 , puisque $\beta^{pps} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. Définissons la décomposition en valeurs singulières de \mathbf{X} comme étant $\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^T$, où les matrices \mathbf{U} , \mathbf{V} et \mathbf{D} diffèrent habituellement de celles de \mathbf{X}_0 , en raison des poids de sondage inhérents.

Le nombre de conditionnement de \mathbf{X} est défini comme étant $\kappa(\mathbf{X}) = \mu_{\max} / \mu_{\min}$, où μ_{\max} et μ_{\min} sont les valeurs singulières maximale et minimale de \mathbf{X} . Le nombre de conditionnement de \mathbf{X} est également différent du nombre des poids de sondage inhérents. Les indices de conditionnement sont définis comme

$$\eta_k = \mu_{\max} / \mu_k, \quad k = 1, \dots, p \tag{6}$$

où μ_k est l'une des valeurs singulières de \mathbf{X} . Les indices de conditionnement et les nombres de conditionnement mis à l'échelle sont les indices de conditionnement et les nombres de conditionnement de la matrice \mathbf{X} mise à l'échelle.

Basé sur les extrêmes du ratio des formes quadratiques (Lin 1984), le nombre de conditionnement $\kappa(\mathbf{X})$ est borné dans l'intervalle :

$$\frac{1}{2} \kappa(\mathbf{X}) \leq \kappa(\mathbf{X}) \leq \frac{W_{\max}^{1/2}}{W_{\min}^{1/2}} \kappa(\mathbf{X}), \tag{7}$$

où W_{\min} et W_{\max} sont les poids de sondage minimal et maximal. Cette expression indique que, si les poids de sondage ne varient pas trop, le nombre de conditionnement dans les MCPPS ressemble à celui dans les MCO. En revanche, pour un échantillon présentant une grande gamme de poids de sondage, le nombre de conditionnement peut être très différent dans les MCPPS et les MCO. Quand le nombre de conditionnement des MCPPS est grand, il se peut que celui des MCO ne le soit pas. Dans le cas d'une dépendance linéaire exacte entre les colonnes de \mathbf{X} , les colonnes de \mathbf{X} seront également linéairement dépendantes. Dans ce cas extrême, au moins une valeur propre de \mathbf{X} sera nulle, et $\kappa(\mathbf{X})$ et $\kappa(\mathbf{X}_0)$ seront toutes deux infinies. Comme dans les MCO, de grandes valeurs de κ ou de η_k , égales ou supérieures à 10, peuvent indiquer qu'il existe des dépendances moyennes à fortes entre deux colonnes ou plus de \mathbf{X} .

\mathbf{W} est la matrice des poids de sondage. La section 3 décrit le cas plus raisonnable.

Dans la décomposition de la variance (3), toutes choses étant égales par ailleurs, une faible valeur singulière μ_j peut donner lieu à une grande composante de $\text{Var}(\beta_j^k)$. Toutefois, si $\nu_{\beta_j^k}$ est petit aussi, $\text{Var}(\beta_j^k)$ peut ne pas être affecté par une petite valeur μ_j . Un cas extrême est celui où $\nu_{\beta_j^k} = 0$. Supposons que les k^{e} et j^{e} colonnes de \mathbf{X} appartiennent à des blocs orthogonaux distincts. Soit $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ avec $\mathbf{X}_1^T \mathbf{X}_2 = \mathbf{0}$ et soit les décompositions en valeurs singulières de \mathbf{X}_1 et \mathbf{X}_2 données, respectivement, par $\mathbf{X}_1 = \mathbf{U}_1 \mathbf{D}_1 \mathbf{V}_1^T$ et $\mathbf{X}_2 = \mathbf{U}_2 \mathbf{D}_2 \mathbf{V}_2^T$. Puisque \mathbf{U}_1 et \mathbf{U}_2 sont les bases orthogonales pour l'espace couvert par les colonnes de \mathbf{X}_1 et de \mathbf{X}_2 , respectivement, $\mathbf{X}_1^T \mathbf{X}_2 = \mathbf{0}$ implique que $\mathbf{U}_1^T \mathbf{U}_2 = \mathbf{0}$ et $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2]$ est à colonnes orthogonales. La décomposition en valeurs singulières de \mathbf{X} est simplement $\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{U}_2^T$ avec

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{bmatrix} \tag{4}$$

et $\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix}$.

Puisqu'au moins un $\nu_{\beta_j^k}$ doit être non nul dans (3), cela implique qu'une forte proportion de toute variance peut être associée à une grande valeur singulière, même en l'absence de colinéarité. L'approche classique consiste à vérifier un indice de conditionnement élevé associé à une forte proportion des variances de deux coefficients ou plus lorsque l'on fait le diagnostic de colinéarité, puisqu'il faut qu'au moins deux colonnes de \mathbf{X} entrent en jeu pour produire une quasi-dépendance. Belsley et coll. (1980) ont proposé de montrer la matrice $\mathbf{\Pi}$ et les indices de conditionnement de \mathbf{X} dans un tableau de décomposition de la variance tel que celui qui suit. Si deux éléments ou plus de la j^{e} ligne de la matrice $\mathbf{\Pi}$ sont relativement grands et que son indice de conditionnement η_j associé est grand également, le signal est que des quasi-dépendances influencent les estimations par la régression.

Indice de conditionnement		Proportions de la variance	
η_1	$\text{Var}_M(\beta_1)$	π_{11}	$\text{Var}_M(\beta_1)$
\vdots	\vdots	\vdots	\vdots
η_p	π_{p1}	π_{p2}	π_{pp}
\vdots	\vdots	\vdots	\vdots
η_2	π_{21}	π_{22}	π_{2p}
\vdots	\vdots	\vdots	\vdots
η_p	π_{p1}	π_{p2}	π_{pp}

sommes et produits non nécessaires, ce qui peut entraîner une erreur de troncation inutile.

Le nombre de conditionnement de \mathbf{X} est défini comme étant $\kappa(\mathbf{X}) = \mu^{\max} / \mu^{\min}$, où μ^{\max} et μ^{\min} sont les valeurs singulières maximale et minimale de \mathbf{X} . Les indices de conditionnement sont définis comme $\eta_k = \mu^{\max} / \mu_k$. Plus μ^{\min} est proche de zéro, plus $\mathbf{X}^T \mathbf{X}$ s'approche d'une matrice singulière. Empiriquement, si une valeur de κ ou η est supérieure à une valeur seuil de, disons, 10 à 30, deux colonnes ou plus de \mathbf{X} présentent des liens moyens à forts. L'occurrence simultanée de plusieurs grandes valeurs de η_k est toujours un indice de l'existence de plus d'une quasi-dépendance.

L'une des questions associées à la DVS est celle de savoir s'il faut centrer les matrices \mathbf{X} autour de leur moyenne. Marquardt (1980) maintient que le centrage des observations élimine le mauvais conditionnement non essentiel. En revanche, Belsley (1984) soutient que le rôle du terme constant dans toute quasi-dépendance sous-jacente. Un cas type est celui de la régression avec variables indicatrices. Par exemple, si le sexe est l'une des variables indépendantes dans une régression et que la plupart des cas étudiés sont des hommes (ou des femmes), la variable indicatrice de sexe peut présenter une forte colinéarité avec l'ordonnée à l'origine. Les discussions consécutives à Belsley (1984) illustrent les divergences d'opinions entre les praticiens (Wood 1984; Snee et Marquardt 1984; Cook 1984). Qui plus est, en analyse par régression linéaire, Wismann, Toutenburg et Shalabh (2007) ont découvert que le choix de la catégorie de référence peut influencer le degré de multicollinéarité avec les variables indicatrices. Dans le présent article, nous ne centrons pas les matrices \mathbf{X} , mais nous illustrons l'effet du choix de la catégorie de référence à la section 4.

Un autre problème associé au nombre de conditionnement est qu'il est affecté par l'échelle des mesures x (Steward 1987). En réduisant l'échelle de toute colonne de \mathbf{X} , le nombre de conditionnement peut être rendu arbitrairement grand. Cette situation est dénommée *conditionnement artificiellement mauvais*. Belsley (1991) propose de modifier l'échelle de chaque colonne de la matrice de plan \mathbf{X} en utilisant la norme euclidienne de chaque colonne avant de calculer le nombre de conditionnement. Cette méthode est implémentée dans SAS et dans le logiciel *perturb* du logiciel statistique R (Hendrickx 2010). Les deux logiciels utilisent comme procédure standard la racine carrée de la moyenne quadratique de chaque colonne pour le changement d'échelle. Le nombre de conditionnement et les indices de conditionnement des matrices \mathbf{X} mises à l'échelle sont appelés *nombre de conditionnement mis à l'échelle* et *indices de conditionnement mis à l'échelle* de la

matrice \mathbf{X} . De même, les proportions découlant de la décomposition de la variance pertinentes pour la matrice \mathbf{X} mise à l'échelle (dont nous discuterons à la section suivante) seront appelées *proportions de décomposition de la variance mises à l'échelle*.

2.3 Méthode de décomposition de la variance

Afin d'évaluer la mesure dans laquelle les quasi-dépendances (c'est-à-dire l'existence d'indices de conditionnement élevés pour \mathbf{X} et $\mathbf{X}^T \mathbf{X}$) dégradent la variance estimée de chaque coefficient de régression, Belsley et coll. (1980) ont réinterprété et étendu les travaux de Silvey (1969) en décomposant la variance d'un coefficient en une somme de termes associés chacun à une valeur singulière. Dans la suite de la présente section, nous examinons les résultats des moindres carrés ordinaire (MCO) sous le modèle $E_M(\mathbf{Y}) = \mathbf{X}\beta$ et $\text{Var}_M(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$ où \mathbf{I}_n est la matrice identité de dimensions $n \times n$. À la section 3, ces résultats seront étendus aux moindres carrés pondérés par les poids de sondage. Rappelons que la matrice de variance-covariance sous le modèle de l'estimateur MCO $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ est $\text{Var}_M(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$. En utilisant la décomposition en valeurs singulières, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, $\text{Var}_M(\hat{\beta})$ peut s'écrire :

$$\text{Var}_M(\hat{\beta}) = \sigma^2 [(\mathbf{U}\mathbf{D}\mathbf{V}^T)^T]^{-1} = \sigma^2 \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T \quad (2)$$

et le k^{e} élément diagonal dans $\text{Var}_M(\hat{\beta})$ est la variance estimée pour le k^{e} coefficient, $\hat{\beta}_k$. En utilisant (2), $\text{Var}_M(\hat{\beta}_k)$ peut s'exprimer :

$$\text{Var}_M(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^J \frac{H_{kj}^2}{\lambda_j^2} \quad (3)$$

où $\mathbf{V} = (\mathbf{V}_j)_{j=1}^J$. Soit $\phi_j = \mathbf{V}_j^T \mathbf{H}_j^T \mathbf{V}_j = \mathbf{V}_j^T \mathbf{H}_j^T \mathbf{V}_j$, où \cdot est le produit (par élé-ment) de Hadamard. Les proportions de décomposition de la variance sont données par $\pi_j = \phi_j / \phi_k$, qui est la proportion de la variance du k^{e} coefficient de régression associée à la j^{e} composante de la matrice de décomposition de la variance par $\Pi = (\pi_j)_{j=1}^J$, où $\hat{\mathbf{Q}}$ est la matrice diagonale contenant les sommes de ligne de $\hat{\mathbf{Q}}$ sur la diagonale principale et des 0 ailleurs.

Si le modèle est donné par $E_M(\mathbf{Y}) = \mathbf{X}\beta$, $\text{Var}_M(\mathbf{Y}) = \sigma^2 \mathbf{W}^{-1}$ et que l'on utilise la méthode des moindres carrés pondérés, $\hat{\beta}_{\text{MCP}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$ et $\text{Var}_M(\hat{\beta}_{\text{MCP}}) = \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$. La décomposition en (3) est vérifiée avec $\mathbf{X} = \mathbf{W}^{1/2} \mathbf{X}$ décomposée comme $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Cependant, dans les applications d'enquête, la matrice de covariance de \mathbf{Y} ne sera virtuellement jamais $\sigma^2 \mathbf{W}^{-1}$ si

(1) $\mathbf{X}\mathbf{v} = \mathbf{0}$, ou $\mathbf{X}^d\mathbf{v} = \mathbf{0}$, $\mathbf{X}^d = \mathbf{v}_1\mathbf{X}^1 + \dots + \mathbf{v}_d\mathbf{X}^d$

Cependant, en pratique, si la matrice de données ne présente pas de colinéarité exacte, mais plutôt certaines quasi-dépendances, il est parfois possible de trouver un ou plusieurs vecteurs non nuls \mathbf{v} tels que $\mathbf{X}\mathbf{v} = \mathbf{a}$ avec $\mathbf{a} \neq \mathbf{0}$, mais proche de $\mathbf{0}$. Ou bien, nous pourrions dire que une quasi-dépendance existe si la longueur du vecteur \mathbf{a} , $\|\mathbf{a}\|$, est petite. Pour normaliser le problème consistant à trouver l'ensemble de vecteurs \mathbf{v} qui rend $\|\mathbf{a}\|$ petite, nous considérons uniquement les vecteurs \mathbf{v} de longueur unitaire, c'est-à-dire tels que $\|\mathbf{v}\| = 1$. Belsley (1991) discute du lien des valeurs propres et des vecteurs propres de $\mathbf{X}^T\mathbf{X}$ avec le vecteur normalisé \mathbf{v} et $\|\mathbf{a}\|$. La longueur minimale $\|\mathbf{a}\|$ est simplement la racine carrée positive de la plus petite valeur propre de $\mathbf{X}^T\mathbf{X}$. Le \mathbf{v} qui produit le vecteur \mathbf{a} de longueur minimale doit être le vecteur propre de $\mathbf{X}^T\mathbf{X}$ qui correspond à la plus petite valeur propre. Comme il est discuté à la section suivante, les valeurs propres et les vecteurs propres de \mathbf{X} sont reliés à ceux de $\mathbf{X}^T\mathbf{X}$ et offrent certains avantages lorsque l'on examine la colinéarité.

2.2 Décomposition en valeurs singulières, nombre de conditionnement et indices de conditionnement

La décomposition en valeurs singulières (DVS) de la matrice \mathbf{X} est très étroitement apparentée au système propre de $\mathbf{X}^T\mathbf{X}$, mais possède des propres avantages. La matrice \mathbf{X} de dimensions $n \times d$ peut être décomposée comme $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, où $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_d$ et $\mathbf{D} = \text{diag}(d_1, \dots, d_p)$ est la matrice diagonale des valeurs singulières (ou valeurs propres) de \mathbf{X} . Ici, les trois composantes de la décomposition sont des matrices très spéciales, possédant des propriétés hautement exploitables : \mathbf{U} est de dimensions $n \times d$ (la même taille que \mathbf{X}) et est à colonnes orthogonales ; \mathbf{V} est de dimensions $d \times d$ et est à colonnes orthogonales ; \mathbf{D} est de dimensions $d \times p$, l'ongnes et lignes orthogonales ; \mathbf{D} est de dimensions $d \times p$, non négative et diagonale. Belsley et coll. (1980) ont estimé que la DVS de \mathbf{X} offrait plusieurs avantages par rapport au système propre de $\mathbf{X}^T\mathbf{X}$, tant sur le plan des usages statistiques que de la complexité des calculs. Pour la prédiction, on se concentre sur \mathbf{X} plutôt que sur la matrice à produit croisé $\mathbf{X}^T\mathbf{X}$, puisque $\mathbf{Y} = \mathbf{X}\mathbf{\beta}$. En outre, les longueurs $\|\mathbf{a}\|$ des combinaisons linéaires (1) de \mathbf{X} qui sont reliées à la colinéarité sont définies convenablement en fonction des racines carrées des valeurs propres de $\mathbf{X}^T\mathbf{X}$, qui sont les valeurs singulières de \mathbf{X} . Un deuxième élément à prendre en considération, étant donné la puissance de calcul actuelle, est que la décomposition en valeurs singulières de \mathbf{X} évite le fardeau de calcul supplémentaire lié à la formation de $\mathbf{X}^T\mathbf{X}$, une opération qui comprend np^2

Nous supposons dans tout l'exposé que les poids de sondage sont construits de façon qu'ils puissent être utilisés pour estimer les totaux de population finie. L'estimateur par les moindres carrés pondérés par les poids de sondage (MCPPS) est donné par

$\hat{\beta}^{pps} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \equiv \mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$,

en supposant que $\mathbf{A} = \mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}$ est inversible. Fuller (2002) décrit les propriétés de cet estimateur. L'estimateur $\hat{\beta}^{pps}$ est modélisé sans biais pour β sous le modèle $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ que $\text{Var}^{pp}(\mathbf{e}) = \sigma^2\mathbf{R}$ soit spécifiée correctement ou non, et est approximativement sans biais sous le plan pour le paramètre de recensement $\mathbf{B}_U = (\mathbf{X}_U^T\mathbf{X}_U)^{-1}\mathbf{X}_U^T\mathbf{Y}_U$, dans la population finie U de N unités. Les valeurs de population finie du vecteur de réponses et de la matrice des variables explicatives sont $\mathbf{Y}_U = (Y_1^p, \dots, Y_N^p)^T$ et $\mathbf{X}_U = (X_1^p, \dots, X_d^p)$ où \mathbf{X}_k est le vecteur de dimension $N \times 1$ des valeurs de la covariable k .

La présentation de l'article est la suivante. À la section 2, nous passons en revue les résultats concernant les nombres de conditionnement et les décompositions de variance pour les moindres carrés ordinaires. À la section 3, nous les étendons afin de les adapter à l'estimation fondée sur des données d'enquête. À la section 4, nous donnons certains exemples numériques des techniques. À la section 5, nous présentons nos conclusions. Dans la plupart des dérivations, nous utilisons des calculs fondés sur un modèle car les formes des variances fondées sur un modèle sont utiles pour comprendre les effets de la colinéarité. Cependant, lorsque nous présentons des décompositions de variance, nous utilisons des estimateurs justifiés à la fois par un modèle et par le plan de sondage.

2. Indices de conditionnement et décompositions de variance dans l'estimation par les moindres carrés ordinaires

À la présente section, nous passons brièvement en revue les techniques de diagnostic de la colinéarité dans l'estimation par les moindres carrés ordinaires (MCO) fondées sur des indices de conditionnement et des décompositions de variance. À la section 3, nous étendrons ces méthodes au cas des données d'enquête complexes.

2.1 Valeurs propres et vecteurs propres de $\mathbf{X}^T\mathbf{X}$

S'il existe une relation de colinéarité exacte (parfaite) dans la matrice de données \mathbf{X} de dimensions $n \times p$, nous pouvons trouver un ensemble de valeurs, $\mathbf{v} = (v_1, \dots, v_p)$, non nulles, tel que

Indices de conditionnement et décompositions des variances pour le diagnostic de la colinéarité dans l'analyse de données d'enquête au moyen de modèles linéaires

Dan Liao et Richard Valliant¹

Résumé

Les colinéarités entre les variables explicatives des modèles de régression linéaire affectent les estimations fondées sur des données d'enquête autant que celles fondées sur des données ne provenant pas d'enquêtes. Les effets indésirables sont des erreurs-types inutilement grandes, des statistiques faussées et des estimations des paramètres de signe illogique. Les diagnostics de colinéarité disponibles ne conviennent généralement pas pour les données d'enquête, parce que les estimateurs de variance qui y sont intégrés ne tiennent pas compte correctement de la stratification, des grappes et des poids de sondage. Dans le présent article, nous élaborons des indices de conditionnement et des décompositions de variance pour diagnostiquer les problèmes de colinéarité dans des données provenant d'enquêtes complexes. Les diagnostics adaptés sont illustrés au moyen de données provenant d'une enquête sur les caractéristiques de l'état de santé.

Mots clés : Diagnostics pour données d'enquête ; multicolinéarité ; décomposition en valeurs singulières ; inflation de la variance.

1. Introduction

Lorsque les variables explicatives d'un modèle de régression sont corréliées entre elles, on parle de colinéarité. Les effets indésirables de cette dernière sont l'obtention d'erreurs-types inutilement grandes, de statistiques faussées, et d'estimations des paramètres de signe illogique ou exagérément sensibles à de faibles variations des valeurs des données. Dans un plan expérimental, il peut être possible de créer des situations où les variables explicatives sont orthogonales les unes par rapport aux autres, mais il n'en va pas de même des données d'observation. Belsley (1991) a souligné que : [traduction] « [...] dans les sciences non expérimentales, ... la colinéarité est une loi naturelle dans l'ensemble de données résultant des opérations incontrôlables du mécanisme de création des données, et est simplement une réalité douloureuse et inévitable. » Dans de nombreuses enquêtes, des données sur l'analyse. Peu d'analystes des données d'enquête échappent au problème de la colinéarité dans l'estimation par la régression, et l'existence de ce problème complique l'explication statistique précise des relations entre les variables explicatives et les réponses.

Alors que de nombreux diagnostics de régression existent pour les données ne provenant pas d'enquêtes, leur nombre est considérablement plus faible pour les données d'enquête. Les quelques articles existants se concentrent sur la détection des points influents et des groupes influents ayant des valeurs de données ou de poids de sondage anormaux. Elliott (2007) a élaboré des méthodes bayésiennes de troncature

des poids des estimateurs par la régression linéaire et par la régression linéaire généralisée sous des plans avec probabilités d'inclusion inégales. Li (2007a, b) et Li et Valliant (2009, 2011) ont étendu une série de techniques diagnostiques classiques à la régression appliquée à des données d'enquête complexes. Leurs articles portent sur les résidus et les effets leviers, plusieurs diagnostics fondés sur la suppression de cas (DFBETA, DFBFATS, DFFIT, DFFITS et distance de Cook) et l'approche pas à pas ascendante (*forward search*). Alors que de nombreuses publications de statistiques appliquées offrent des suggestions et des lignes directrices précieuses pour aider les analystes des données à diagnostiquer la présence de colinéarité (par exemple, Belsley, Kuh et Welsch 1980; Belsley 1991; Farrar et Clauber 1967; Fox 1986; Theil 1971), presque aucun de ces travaux de recherche ne traite des diagnostics de colinéarité lorsque les modèles sont ajustés en se servant de données d'enquête. Un article antérieur portant sur les problèmes de colinéarité dans le contexte des enquêtes est celui de Liao et Valliant (2012) qui ont adapté des facteurs d'inflation de la variance pour des modèles linéaires ajustés à des données d'enquête.

Supposons que le modèle structurel sous-jacent dans la superpopulation est $Y = X\beta + e$. La matrice X est une matrice de dimensions $n \times p$ de variables explicatives où n est la taille de l'échantillon ; β est un vecteur de dimension $p \times 1$ de paramètres. Les termes d'erreur du modèle ont une structure de variance générale $e \sim (0, \sigma^2 R)$ où σ^2 est une constante inconnue et R est une matrice de covariance de dimensions $n \times n$ inconnue. Définitions W comme étant la matrice diagonale des poids de sondage.

1. Dan Liao, RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005. Courriel : dliao@rti.org ; Richard Valliant, University of Michigan et University of Maryland, Joint Program in Survey Methodology, 1218 LeFrak Hall, College Park, MD, 20742.

- Cho, M., Eitinge, J., Gershunskaya, J. et Huff, L. (2002). Evaluation of generalized variance function estimators for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 534-539.
- Fay, R., et Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Gershunskaya, J., et Lahiri, P. (2005). Variance estimation for domains in the U.S. current employment statistics program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3044-3051.
- Ghosh, M., et Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 54-76.
- Hall, P., et Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested-error regression models. *Annals of Statistics*, 34, 1733-1750.
- Huff, L., Eitinge, J. et Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1519-1524.
- Hwang, J., Qiu, J. et Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both mean and variances. *Journal of the Royal Statistical Society*, B, 71, 265-285.
- Joshi, V. (1969). Admissibility of the usual confidence sets for the *Mathematical Statistics*, 40, 1042-1067.
- Maples, J., Bell, W. et Huang, E. (2009). Small area variance modelling with application to county poverty estimates from the american community survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 5056-5067.
- Choi, M., et Bell, W. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 160-165.
- Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Revue Internationale de Statistique*, 70, 125-143.
- Prasad, N., et Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Qiu, J., et Hwang, J. (2007). Sharp simultaneous intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, 63, 767-776.
- Rao, J. (2003). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 145-169.
- Rivest, L.-P., et Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*.
- Robert, C., et Casella, G. (2004). *Monte Carlo Statistical Methods* (Deuxième édition).
- Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82, 499-508.
- Wang, J., et Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *Journal of the American Statistical Association*, 98, 716-723.
- You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances au d'échantillonnage. *Techniques d'enquête*, 32, 1, 107-114.

$$(n_i - 1)S_i^2 \Big| \sigma_i^2 \sim \chi_{n_i-1}^2 \quad (C.3)$$

$$\sigma_i^2 \sim \text{Inverse} - \text{Gamma}(a, b), \quad (C.4)$$

qui peut être considéré comme une distribution à échelle possédant $n_i + 2a$ degrés de liberté et le paramètre d'échelle $\delta^* \lambda / (1 + \lambda)$ avec $\delta^{*2} = \delta^2 / (n_i + 2a)$. D'où,

$$\pi(\theta_i | X_i, S_i^2, \mathbf{B}) = \frac{\delta^* \Gamma(n_i/2 + a) \int (n_i + 2a) \lambda \pi}{\Gamma((n_i + 1)/2 + a) (1 + \lambda)} \left\{ 1 + \frac{(n_i + 2a) \delta^{*2} \lambda / (1 + \lambda)}{(\theta_i - \mu_i)^2} \right\}^{-\frac{1}{2}(n_i + 2a + 1/2)}$$

indépendamment pour $i = 1, 2, \dots, n$. Notons que, dans la formule susmentionnée, il est supposé que la variance conditionnelle de θ_i est proportionnelle à σ_i^2 , tandis que la variance marginale est constante (en éliminant σ_i^2 par intégration en utilisant (C.4)). Dans (1) et (2), la variance de θ_i est une constante, τ^2 , indépendante de σ_i^2 , et il n'existe pour θ_i aucune structure conditionnelle dépendant de σ_i^2 . L'ensemble de tous les paramètres inconnus dans le modèle hiérarchique courant est $\mathbf{B} = (a, b, \beta, \lambda)$. La procédure d'inférence pour ce modèle est donnée ci-après. Le modèle repose essentiellement sur l'hypothèse que les effets réels de petit domaine ne sont pas identiquement distribués, même après avoir éliminé les variations connues.

C.1 Méthodologie d'inférence

En reparamétrisant la variance comme dans (C.2), on obtient certaines simplifications analytiques pour dériver les lois a posteriori de θ_i et σ_i^2 sachant X_i, S_i^2 et \mathbf{B} . Nous avons

$$\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{B}) = GI \left(\frac{n_i}{2} + a, \left[\frac{(n_i - 1)S_i^2}{2} + \frac{(X_i - \mathbf{Z}^T \beta)^2}{2(1 + \lambda)} + \frac{1}{b} \right]^{-1} \right)$$

où $GI(a, b)$ représente la loi Gamma inverse dont les paramètres de forme et d'échelle sont a et b , respectivement. Sachant \mathbf{B} et σ_i^2 , la distribution conditionnelle de θ_i est

$$\pi(\theta_i | X_i, \sigma_i^2, \mathbf{B}) = \text{Normale} \left(\mathbf{Z}_i^T \beta, \frac{\lambda \sigma_i^2}{1 + \lambda} \right).$$

En éliminant σ_i^2 par intégration, on obtient la distribution conditionnelle de θ_i sachant X_i, S_i^2 et \mathbf{B} ,

$$\pi(\theta_i | X_i, S_i^2, \mathbf{B})$$

$$= \int_{-\infty}^{\infty} \pi(\theta_i | X_i, \sigma_i^2, \mathbf{B}) \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{B}) d\sigma_i^2 = \infty \left\{ \frac{1 + \lambda}{(1 + \lambda)} (\theta_i - \mathbf{Z}_i^T \beta)^2 + \frac{2}{\delta^2} \right\}^{-\frac{1}{2}(n_i + 2a + 1/2)}$$

où $\delta^2 = (n_i - 1) S_i^2 + (X_i - \mathbf{Z}_i^T \beta)^2 / (1 + \lambda) + 2/b$. Nous pouvons réécrire (C.5) sous la forme

Bibliographie

- Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 95, 28-36.
- Bell, W. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. Rapport technique du U.S. Census Bureau.
- Casella, G. et Hwang, J. (1991). Evaluating confidence sets using loss functions. *Statistica Sinica*, 1, 159-173.
- Chatterjee, S., Lahiri, P. et Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics*, 36, 1221-1245.

θ_i , sachant les données et \mathbf{B} sont

$$\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) = \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 | \mathbf{Z}_i, \mathbf{B}) d\theta_i \propto \frac{(\sigma_i^2)^{(n_i-1)/2 + a + 1} (\sigma_i^2 + \tau_i^2)^{1/2}}{1} \exp \left[-\frac{2}{(\sigma_i^2 + \tau_i^2)} \left(X_i - \mathbf{Z}_i^T \boldsymbol{\beta} \right)^2 - \frac{1}{2} (n_i - 1) S_i^2 + \frac{b}{1} \right] \left(\frac{\sigma_i^2}{1} \right),$$

$$\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) = \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 | \mathbf{Z}_i, \mathbf{B}) d\sigma_i^2 \propto \exp \left\{ -\frac{2\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \right\} \psi_i \left(-\frac{\tau_i^2}{n_i + a} \right)$$

où ψ_i est définie dans l'équation (4).

B. Détails de l'algorithme EM

La maximisation de $\tilde{Q}(\mathbf{B} | \mathbf{B}^{(t-1)})$ est effectuée en posant que les dérivées partielles par rapport à \mathbf{B} sont nulles, c'est-à-dire

$$\frac{\partial \tilde{Q}(\mathbf{B} | \mathbf{B}^{(t-1)})}{\partial \mathbf{B}} = 0. \quad (\text{B.1})$$

Partant de l'expression de $\tilde{Q}(\mathbf{B} | \mathbf{B}^{(t-1)})$ dans le corps du

texte, nous obtenons des expressions explicites pour les dérivées partielles par rapport à chaque composante de \mathbf{B} .

La dérivée partielle correspondant à $\boldsymbol{\beta}$ est

$$\frac{\partial \tilde{Q}(\mathbf{B} | \mathbf{B}^{(t-1)})}{\partial \boldsymbol{\beta}} = \frac{\sum_{i=1}^n \int \mathbf{Z}_i^T \left(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta} \right) \exp \left(-\frac{\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \right) \psi_i \left(-\frac{\tau_i^2}{n_i + a} \right) d\theta_i}{\sum_{i=1}^n \int \mathbf{Z}_i^T \left(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta} \right) \exp \left(-\frac{\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \right) \psi_i \left(-\frac{\tau_i^2}{n_i + a} \right) d\theta_i} = \sum_{i=1}^n \mathbf{E} \left\{ \mathbf{Z}_i^T \left(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta} \right) \exp \left(-\frac{\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \right) \right\}$$

où l'espérance est calculée par rapport à la distribution conditionnelle de θ_i , $\pi(\theta_i | X_i, S_i^2, \mathbf{B})$. L'expression de la dérivée partielle correspondant à τ_i^2 est :

$$\frac{\partial \tilde{Q}(\mathbf{B} | \mathbf{B}^{(t-1)})}{\partial \tau_i^2} = -\frac{\sum_{i=1}^n \frac{2\tau_i^2}{n_i} - \sum_{i=1}^n \int \frac{2\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \exp \left(-\frac{\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \right) \psi_i \left(-\frac{\tau_i^2}{n_i + a} \right) d\theta_i}{\sum_{i=1}^n \int \frac{2\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \exp \left(-\frac{\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \right) \psi_i \left(-\frac{\tau_i^2}{n_i + a} \right) d\theta_i} = -\frac{\sum_{i=1}^n \frac{2\tau_i^2}{n_i} + \frac{2\tau_i^2}{n_i} \sum_{i=1}^n \int \frac{2\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \exp \left(-\frac{\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \right) \psi_i \left(-\frac{\tau_i^2}{n_i + a} \right) d\theta_i}{\sum_{i=1}^n \int \frac{2\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \exp \left(-\frac{\tau_i^2}{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2} \right) \psi_i \left(-\frac{\tau_i^2}{n_i + a} \right) d\theta_i}.$$

C. Une autre formulation du modèle d'estimation sur petits domaines

Il est possible de réduire la largeur de l'intervalle de confiance $\tilde{C}(\mathbf{B})$ en se fondant pour l'estimation sur petits domaines sur un autre modèle hiérarchique qui présente une certaine élégance mathématique. Dans (19), le terme constant $n_i + 2a + 2$ devient $n_i + 2a$ dans cette autre formulation du modèle. Le modèle est donné par

$$X_i | \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \quad (\text{C.1})$$

$$\theta_i | \sigma_i^2 \sim N(\mathbf{Z}_i^T \boldsymbol{\beta}, \lambda \sigma_i^2), \quad (\text{C.2})$$

et

$$S_{bb} = \sum_{i=1}^n \left(\frac{b^2}{a} - (n_i + 2a) \frac{1}{a} E \left(\frac{\psi_i}{1} \right) + \left(\frac{\psi_i}{1} + a \right) \frac{b^2}{2} \text{Var} \left(\frac{\psi_i}{1} \right) \right) \quad (\text{B.2})$$

avec $S^{ba} = S^{ab}$. À la n^e étape, les mises à jour de a et b sont données par

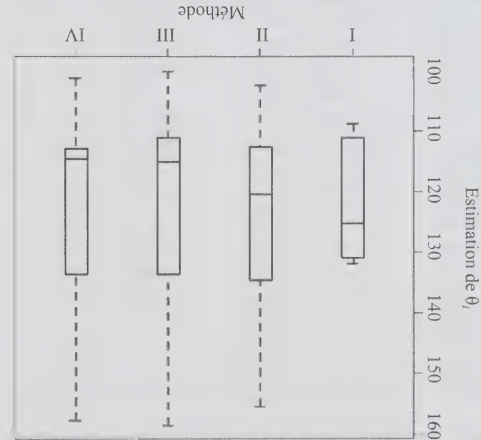
$$\begin{bmatrix} a^{(n)} \\ b^{(n)} \end{bmatrix} = \begin{bmatrix} a^{(n-1)} \\ b^{(n-1)} \end{bmatrix} - \begin{bmatrix} S_{aa}^{(n-1)} & S_{ab}^{(n-1)} \\ S_{ab}^{(n-1)} & S_{bb}^{(n-1)} \end{bmatrix}^{-1} \begin{bmatrix} S_a^{(n-1)} \\ S_b^{(n-1)} \end{bmatrix}, \quad (\text{B.3})$$

où l'indice supérieur $(n-1)$ sur S^{aa} , S^{ab} , S^{ba} , S_a et S_b désigne ces quantités évaluées aux valeurs qu'avait a et b à la $(n-1)^e$ itération. Lorsque la procédure de Newton-Raphson converge, les valeurs de a et b à la t^e étape de l'algorithme EM sont fixées à $a^{(t)} = a^{(\infty)}$ et $b^{(t)} = b^{(\infty)}$.

Le présent article décrit la modélisation conjointe au niveau du domaine des moyennes et des variances pour l'estimation sur petits domaines. Il montre que les estimateurs sur petits domaines résultants sont plus efficaces que les estimateurs classiques obtenus en utilisant les modèles de Fay-Herriot qui ne réécussent que les moyennes. Bien que notre modèle soit le même que celui pris en considération dans Hwang et coll. (2009), notre méthode d'estimation diffère à deux égards, en ce qui concerne la détermination du paramètre de mise au point k et l'utilisation de $\pi(\sigma_i^2 | X_i, S_i^2, Z_i)$ (qui dépend additionnellement de X_i), au lieu de $\pi(\sigma_i^2 | S_i^2, Z_i)$ pour construire la distribution conditionnelle des paramètres θ_i de petit domaine. Nous avons démontré les propriétés de robustesse du modèle quand l'hypothèse que σ_i^2 est issue d'une loi Gamma inverse est violée. L'emprunt de l'information X_i pour estimer σ_i^2 ainsi que la robustesse à l'élitisation de la loi a priori démontre la supériorité de la méthode que nous proposons. Les valeurs des paramètres choisis dans l'étude par simulation diffèrent de celles utilisées dans l'analyse des données réelles. Cette dernière est présentée ici simplement en guise d'illustration. Notre objectif principal était d'élaborer la méthodologie de modélisation de la moyenne et de la variance, et de la comparer à certaines méthodes étroites ment appariées afin de montrer son efficacité. C'est pour-quoi nous avons choisi de configurer les paramètres dans la simulation de la même façon que dans l'article traitant de

7. Conclusion

Figure 2 Boîtes à moustaches des estimations du nombre d'hectares consacrés au maïs pour chaque comté. (I) à (IV) sont les quatre méthodes correspondant à la figure 1



Les auteurs remercient deux examinateurs et le rédacteur associé de leurs commentaires constructifs qui leur ont permis d'améliorer considérablement l'article. L'étude a été financée en partie par les subventions SES 0961649, 0961618 et DMS 1106450 de la NSF.

Remerciements

L'estimation sur petits domaines bien connu de Wang et Fuller (2003). L'obtention d'estimateurs améliorés de la variance d'échantillonnage est un produit secondaire de l'approche proposée. Nous avons fourni une technique d'estimation novatrice, qui est justifiée théoriquement et facile à utiliser. En ce qui concerne les calculs, la méthode est beaucoup plus simple que certaines méthodes concurrentes telles que les procédures MCMC bayésiennes ou les méthodes de ré-échantillonnage bootstrap. Notre méthode ne requiert qu'un seul échantillonnage à partir de la loi a posteriori durant l'estimation des paramètres du modèle, et les valeurs échantillonnées peuvent être utilisées par la suite à toute autre fin. Le logiciel peut être obtenu sur demande auprès des auteurs.

A. Obtention des distributions conditionnelles

Des équations (1) et (2) il découle que la distribution conjointe conditionnelle de $\{X_i, S_i^2, \theta_i, \sigma_i^2\}$, $\pi(X_i, S_i^2, \theta_i, \sigma_i^2 | a, b, \beta, \tau^2)$ est

$$\pi(X_i, S_i^2, \theta_i, \sigma_i^2 | Z_i, B) = \frac{1}{1} \exp\left\{-\frac{2\pi\sigma_i^2}{(X_i - \theta_i)^2}\right\} \frac{\Gamma\left(\frac{n_i - 1}{2}\right)}{1} \times \left\{\frac{2\sigma_i^2}{(n_i - 1)S_i^2}\right\} \exp\left\{-\frac{2\sigma_i^2}{(n_i - 1)S_i^2}\right\} \times \left\{\frac{\sigma_i^2}{n_i - 1}\right\} \frac{1}{2\pi\tau^2} \exp\left\{-\frac{(\theta_i - Z_i'\beta)^2}{2\tau^2}\right\} \times \left\{\frac{\sigma_i^2}{n_i - 1}\right\} \frac{1}{\Gamma(a)b^a} \exp\left\{-\frac{b\sigma_i^2}{1}\right\} \times \left[\frac{2}{(n_i - 1)S_i^2} + \frac{1}{1}\right] \frac{b}{1} \left\{\frac{\sigma_i^2}{1}\right\} \times \left\{\frac{\Gamma(a)b^a}{1}\right\} \frac{1}{\Gamma\left(\frac{1}{2}\right)} \frac{1}{1} \frac{\Gamma(a)b^a}{1}$$

Annexe

Tableau 7 Résultats de l'analyse des données sur le maïs. Ici, IC et LIC représentent l'intervalle de confiance et la longueur de l'intervalle de confiance, respectivement

Comté	θ_i	IC	LIC	θ_i	IC	LIC
Franklin	131,8106	104,085 ; 159,372	55,287	155,4338	124,151 ; 193,094	68,943
Pocahontas	108,7305	80,900 ; 136,436	55,536	102,3682	-38,973 ; 244,019	282,993
Winnebago	109,0559	81,430 ; 136,646	55,216	115,9093	-53,768 ; 279,314	333,083
Wright	131,6113	103,736 ; 159,564	55,828	131,0674	8,330 ; 280,263	271,932
Webster	113,1484	92,805 ; 133,348	40,543	109,4795	32,514 ; 202,675	170,161
Hancock	129,4279	111,781 ; 147,193	35,412	124,1028	56,750 ; 162,013	105,262
Kossuth	121,0071	103,451 ; 138,626	35,175	116,7147	68,049 ; 152,454	84,405
Hardin	130,2520	112,373 ; 148,114	35,741	137,7983	51,734 ; 188,373	136,638
Franklin	158,4677	128,564 ; 188,370	59,805	157,7383	146,999 ; 168,477	21,478
Pocahontas	100,1276	-44,039 ; 244,295	288,334	101,1661	19,444 ; 182,887	163,442
Winnebago	114,1473	0,065 ; 228,228	228,163	113,7746	56,263 ; 171,286	115,022
Wright	140,3717	-24,119 ; 304,862	328,982	143,2244	41,559 ; 244,889	203,330
Webster	115,7865	50,297 ; 181,275	130,978	115,2224	75,124 ; 155,320	80,196
Hancock	111,3087	66,213 ; 156,403	90,189	113,1766	83,691 ; 142,661	58,970
Kossuth	126,6093	74,366 ; 147,550	73,184	112,3239	89,520 ; 135,127	45,607
Hardin		40,040 ; 213,178	173,137	123,9049	54,607 ; 193,202	138,594
III : Hwang et coll. (2009)						
Franklin	158,4677	128,564 ; 188,370	59,805	157,7383	146,999 ; 168,477	21,478
Pocahontas	100,1276	-44,039 ; 244,295	288,334	101,1661	19,444 ; 182,887	163,442
Winnebago	114,1473	0,065 ; 228,228	228,163	113,7746	56,263 ; 171,286	115,022
Wright	140,3717	-24,119 ; 304,862	328,982	143,2244	41,559 ; 244,889	203,330
Webster	115,7865	50,297 ; 181,275	130,978	115,2224	75,124 ; 155,320	80,196
Hancock	111,3087	66,213 ; 156,403	90,189	113,1766	83,691 ; 142,661	58,970
Kossuth	126,6093	74,366 ; 147,550	73,184	112,3239	89,520 ; 135,127	45,607
Hardin		40,040 ; 213,178	173,137	123,9049	54,607 ; 193,202	138,594
IV : Qiu et Hwang (2007)						
Franklin	158,4677	128,564 ; 188,370	59,805	157,7383	146,999 ; 168,477	21,478
Pocahontas	100,1276	-44,039 ; 244,295	288,334	101,1661	19,444 ; 182,887	163,442
Winnebago	114,1473	0,065 ; 228,228	228,163	113,7746	56,263 ; 171,286	115,022
Wright	140,3717	-24,119 ; 304,862	328,982	143,2244	41,559 ; 244,889	203,330
Webster	115,7865	50,297 ; 181,275	130,978	115,2224	75,124 ; 155,320	80,196
Hancock	111,3087	66,213 ; 156,403	90,189	113,1766	83,691 ; 142,661	58,970
Kossuth	126,6093	74,366 ; 147,550	73,184	112,3239	89,520 ; 135,127	45,607
Hardin		40,040 ; 213,178	173,137	123,9049	54,607 ; 193,202	138,594

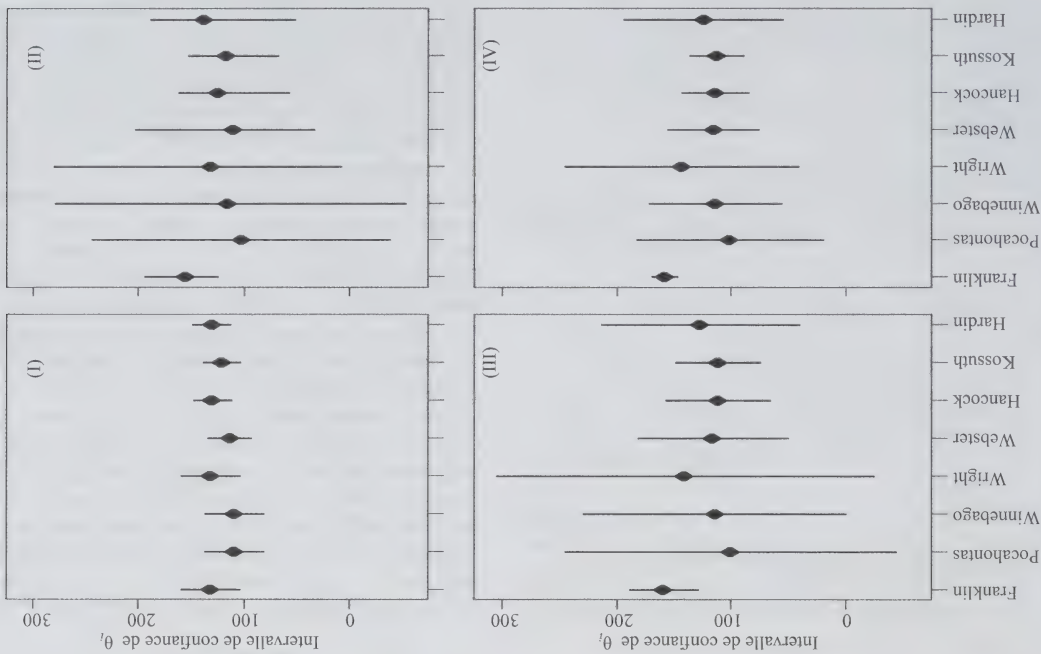


Figure 1 Estimation du nombre d'hectares consacrés au maïs. Pour chaque comté, la droite horizontale donne l'intervalle de confiance de θ_i , avec θ_i marqué par le cercle, pour (I) la méthode proposée, (II) Wang et Fuller (2003), (III) Hwang et coll. (2009) et (IV) Qiu et Hwang (2007)

Tableau 6
Données sur le maïs provenant de You et Chapman (2006)

Comité	n_i	X_i	Z_{1i}	Z_{2i}	$\sqrt{S_i^2}$
Franklin	3	158,623	318,21	188,06	5,704
Pocahontas	3	102,523	257,17	247,13	43,406
Winnebago	3	112,773	291,77	185,37	30,547
Wright	3	144,297	301,26	221,36	53,999
Webster	4	117,595	262,17	247,09	21,298
Hancock	5	109,382	314,28	198,66	15,661
Kossuth	5	110,252	298,65	204,61	12,112
Hardin	5	120,054	325,99	177,05	36,807

Les estimations de B sont les suivantes : $a = 1,707$, $b = 0,00135$, $\tau^2 = 90,58$ et $\beta = (-186,0; 0,7505; 0,4100)$. La moyenne a priori estimée de $1/\sigma_i^2$ qui est la moyenne de la loi Gamma dont les paramètres sont a et b , est $ab = 0,002295$ dont la racine carrée est $0,048$ (notons que $1/0,048 = 20,85$, valeur en harmonie avec l'intervalle de variation des écarts-types d'échantillon s'étendant de $5,704$ à $53,999$). Les estimations sur petits domaines et leurs intervalles de confiance sont résumés au tableau 7 et à la figure 1. Les estimations ponctuelles produites par les quatre méthodes sont comparables : les mesures sommatives comprenant la moyenne, la médiane et l'étendue des estimations des paramètres de petit domaine pour les méthodes I, II, III et IV sont $(121,9; 124,1; 122,2; 122,6)$, $(125,2; 120,4; 115,0; 114,5)$ et $(23,1; 53,0; 58,4; 56,6)$, respectivement. Les distributions de θ_i (représentées graphiquement en prenant en considération tous les i) sont résumées à la figure 2 qui révèle une différence significative de variabilité. La méthode I est celle dont la variabilité est la plus faible et qui est donc la meilleure en ce sens. En outre, le lissage des variances d'échantillonnage a de fortes répercussions sur la mesure de l'incertitude et donc de l'estimation de l'intervalle. La méthode proposée donne l'intervalle de confiance le plus court, en moyenne, comparativement à toutes les autres méthodes. Les méthodes II et III donnent des intervalles dont la borne inférieure est négative, ce qui paraît irréaliste, car la moyenne directe des superficies consacrées à la culture du maïs est positive et grande pour les 12 comités [les intervalles de confiance bruts $(x_i \pm t_{0,025} S_i)$ ne contiennent non plus de valeur nulle pour aucun des domaines]. Il n'existe aucun soutien théorique pour les intervalles de confiance de la méthode II. Les méthodes II et III produisent des intervalles de confiance plus larges quand la variance d'échantillonnage est élevée. Par exemple, la taille d'échantillon pour les comités de Franklin et de Pocahontas est de $5,704$ et

43,406, respectivement. Alors que les intervalles de confiance sont comparables sous la méthode I, ils sont très différents sous les méthodes II et III. II en est ainsi parce que, même si ces méthodes tiennent compte de l'incertitude dans les estimations de la variance d'échantillonnage, comme le lissage n'a pas été effectué en utilisant l'information provenant des estimations directes d'après l'enquête, les estimations de la variance d'échantillonnage sous-jacentes demeurent très variables (à cause de la petite taille d'échantillon). En fait, la variance de l'estimateur de variance (des estimations ponctuelles) est plus grande que celle obtenue lorsque l'on applique la méthode I. Cela est aussi confirmé par le fait que les écarts-types intuitifs des estimations sur petits domaines « lissés » (un quart de l'intervalle) sont plus faibles et moins variables sous la méthode I que sous les autres méthodes. Une autre caractéristique de notre méthode qui mérite d'être soulignée est que les largeurs des intervalles sont comparables pour les comités pour lesquels la taille d'échantillon est la même. Cela pourrait être une indication que l'on obtient des estimateurs équivalents pour des tailles d'échantillon équivalentes.

Choix du modèle : Afin de choisir le modèle le mieux ajusté, nous avons utilisé le critère d'information bayésien (BIC pour *Bayesian Information Criteria*) qui tient compte à la fois de la vraisemblance et de la complexité des modèles ajustés. Nous avons calculé le BIC pour les modèles utilisés dans les méthodes I et III (Hwang et coll. 2009). Ces deux modèles comprennent le même nombre de paramètres et ne diffèrent que par la façon dont ces paramètres sont estimés. Le BIC du modèle est égal à $210,025$ pour la méthode I et à $227,372$ pour la méthode III, ce qui témoigne de la supériorité de notre méthode. Nous n'avons pas pu calculer le BIC pour le modèle de Wang et Fuller (2003), car ils n'ont utilisé aucune fonction de vraisemblance explicite.

que les autres méthodes. La méthode IV a produit des intervalles de longueur comparable à ceux des autres méthodes dans tous les cas sauf quand σ_1^2 était élevé, auquel cas les longueurs étaient considérablement plus grandes. L'intervalle de confiance proposée dans Qiu et Hwang (2007) n'a pas de bonnes propriétés en échantillon fini, particulièrement pour les petites valeurs de τ^2 . Afin d'éviter un faible taux de couverture, ils ont proposé de tronquer $M_0 = \tau^2/(\tau^2 + \sigma_1^2)$ à l'aide d'un nombre positif $M_1 = 1 - \bar{Q}_\alpha/(v - 2)$ pour σ_1^2 connu, où \bar{Q}_α est le α^e quantile d'une distribution du khi-carré à v degrés de liberté. Quand le ratio de la variance d'échantillonnage à la variance du modèle, σ_1^2/τ^2 , est élevé, M_1 a tendance à être plus grand que M_0 , ce qui donne le taux nominal de couverture, mais avec de plus grandes longueurs d'intervalle. Par exemple, dans le cas où $(\sigma_1^2, \tau^2) = (16, 0,25)$, la LMIC est de 11,13 pour la méthode IV, alors qu'elle est seulement de 2,78 et 4,56 pour les méthodes I et II, respectivement.

5.2 Étude de la robustesse

Afin d'étudier la robustesse de la méthode proposée aux écarts par rapport à l'hypothèse de normalité des erreurs, nous avons procédé à l'étude par simulation qui suit. Les données ont été générées comme précédemment, mais en tirant les e_{ij} d'une loi exponentielle double (loi de Laplace) et d'une loi uniforme. Les estimateurs des méthodes II et III ont eu peu d'effet. Cela pourrait tenir au fait que, dans ces méthodes, l'estimation des paramètres du modèle se fait par la méthode des moments. La méthode IV a produit de plus grandes valeurs du biais relatif de l'EQM et de la LMIC, et une plus faible probabilité de couverture. L'EQM est systématiquement plus faible pour la méthode I que pour la méthode II. Quand $\tau^2 = 0,25$ et I, la LMIC est plus petite pour la méthode I que pour la méthode II pour la méthode II. (n = 36,

Pour illustrer notre méthodologie, nous choisissons un exemple très souvent étudié. Le jeu de données, qui provient du U.S. Department of Agriculture, a été analysé pour la première fois par Battese (1988). Il s'agit de données sur les productions de maïs et de soja dans 12 comtés de l'Iowa. Les tailles d'échantillon pour ces domaines sont faibles, variant de 1 à 5. Faute d'espace, nous considérons uniquement le cas du maïs. Pour les modèles proposés, il faut nécessairement que l'on ait des tailles d'échantillon $n_j > 1$. Par conséquent, nous avons utilisé des données modifiées tirées de You et Chapman (2006) avec $n_j \geq 2$. Les nombres déclarés d'hectares consacrés à la culture du maïs (X_i), qui sont les estimations directes par sondage, sont présentés au tableau 6. Ce tableau donne aussi les variances d'échantillonnage qui sont calculées d'après les données originales sous l'hypothèse d'un échantillonnage aléatoire simple. L'écart-type d'échantillon varie fortement, de 5,704 à 53,999 (le coefficient de variation varie de 0,036 à 0,423). Deux covariables sont considérées dans le tableau 6 : Z_{1j} , le nombre moyen de pixels correspondant à du maïs et Z_{2j} , le nombre moyen de pixels correspondant à du soja, provenant des données de satellite LANDSAT.

6. Analyse de données réelles

m = 9), mais le résultat inverse s'observe quand (n = 180, m = 18). Pour ce qui est de la PC, la méthode II produit une certaine sous-couverture (taux le plus faible égal à 80 %). Par contre, la méthode I ne produit aucune sous-couverture. Faute d'espace, nous présentons uniquement les résultats pour les paramètres a, b, β et τ^2 sous les erreurs laplaciennes (tableau 5).

Tableau 5
Résultats des simulations pour les paramètres du modèle, a (panneau supérieur gauche), b (panneau supérieur droit), β (panneau inférieur gauche) et τ^2 (panneau inférieur droit) quand les erreurs suivent une loi de Laplace. Ici, E.-T. représente l'écart-type sur 200 répétitions. Nous avons pris $\beta = 10$ et $\tau^2 = 0,25$, 1 et 4

$n = 36, m = 9$				$n = 180, m = 18$					
τ^2	Moyenne	E.-T.		τ^2	Moyenne	E.-T.			
0,25	0,9624	0,1632	a	0,25	0,5793	0,1733	b		
1	0,9628	0,1657		1	0,5816	0,1777			
4	0,9689	0,1694		4	0,5758	0,1796			
0,25	0,9736	0,3775	β	0,25	0,2696	0,0882	τ^2		
1	0,9753	0,3709		1	1,0508	0,2501			
4	0,9736	0,4835		4	3,9624	1,1719			
0,25	0,9736	0,0074		0,25	0,5279	0,0501			
1	0,9753	0,0668		1	0,5275	0,0503			
4	0,9736	0,4201		4	0,5263	0,0503			

Tableau 4
Résultats des simulations pour la prédiction quand $\tau^2 = 4$. Ici, EQM, LMIC et PC représentent l'erreur quadratique moyenne, la longueur moyenne de l'intervalle de confiance et la probabilité de couverture de l'intervalle de confiance, respectivement

$n = 36, m = 9$				$n = 180, m = 18$			
Méthode		Méthode		Méthode		Méthode	
BIAS	σ_i^2	BIAS	σ_i^2	BIAS	σ_i^2	BIAS	σ_i^2
I	-0.0024	0.0248	0.0229	0.0180	-0.0084	-0.0098	-0.0122
4	-0.0343	-0.0310	-0.0210	-0.0340	-0.0110	-0.0092	-0.0174
16	-0.0147	0.0702	0.0467	0.0016	0.0024	-0.0059	0.0012
I	0.8822	0.8590	0.8579	1.0559	0.8359	0.8180	0.8541
4	2.0577	2.2900	2.1818	2.2422	2.0424	2.1000	2.0935
16	3.4516	3.7600	3.9267	3.8981	3.3153	3.3500	3.3939
I	4.6318	4.5369	3.7677	4.0256	3.5234	3.9626	3.7499
4	6.2015	10.9093	7.0376	6.4314	5.9000	9.0913	6.2217
16	7.7221	18.0039	9.6718	11.3341	7.4430	14.6665	8.3908
I	0.9791	0.9670	0.9733	0.9029	0.9674	0.9570	0.9600
4	0.9556	0.9670	0.9725	0.9496	0.9592	0.9610	0.9633
16	0.9510	0.9670	0.9796	0.9858	0.9573	0.9650	0.9718

Comparaisons des biais : Dans la plupart des cas, les biais des quatre méthodes sont comparables. Il n'existe aucune preuve manifeste d'écarts significatifs entre elles pour ce qui est du biais. Une forte variance d'échantillonnage donne plus de poids à la moyenne de population par construction, ce qui rend l'estimateur plus proche de la moyenne au deuxième niveau. Par ailleurs, les méthodes I à III comprennent l'utilisation d'estimateurs à rétrécisseur des variances d'échantillonnage qui seraient donc inférieurs au maximum de l'ensemble des variances d'échantillonnage. Donc, les méthodes I à III ont tendance à présenter un biais un peu plus important. Cependant, en raison du rétrécissement des variances d'échantillonnage, on peut s'attendre à une amélioration de la variance des estimateurs qui, à son tour, réduit l'EQM. Parmi les méthodes I à III, la méthode I a donné de meilleurs résultats que les méthodes II et III, à donner les propriétés étaient assez semblables. Le gain maximal en utilisant la méthode I au lieu de la méthode II est de 99 %.

Comparaison des EQM : En ce qui concerne l'EQM, la méthode I a donné systématiquement de meilleurs résultats que les trois autres dans tous les cas, sauf quand le ratio de σ_1^2 à τ^2 était le plus faible ($\sigma_1^2 = 1 / (\tau^2 = 4) = 0.25$). Dans ce cas, la variance entre les petits domaines (variance du modèle) est beaucoup plus grande que la variance dans les domaines (variance d'échantillonnage). Lorsque notre méthode est utilisée pour estimer θ_j , l'information « empruntée » à d'autres domaines peut mal orienter l'estimation : la moyenne estimée de la loi Gamma pour σ_j^2 provient du deuxième niveau de (2) est ab , qui est égale à 0.44 environ pour les deux combinaisons (m, n) correspondant à

(9, 36) et (18, 180) (la valeur réelle est $ab = 0.4$). Donc, $E(\sigma_j^2 | X_j, S_j^2, B)$ est significativement plus petite que 1 en raison du rapprochement vers la moyenne pour le groupe pour lequel la valeur réelle est $\sigma_j^2 = 1$. En outre, puisque σ_j^2 est plus faible que τ^2 , le poids de X_j devrait être beaucoup plus élevé comparativement à β_j , la moyenne globale. Cependant, étant donné la sous-estimation de σ_j^2 dans ce cas, l'estimateur résultant donne moins de poids à X_j , ce qui donne lieu à une EQM plus grande. Cependant, cette sous-estimation diminue pour les grandes tailles d'échantillon en raison de la cohérence des estimateurs de Bayes. Ce fait s'observe effectivement quand la taille d'échantillon passe de $n = 36$ à $n = 180$ pour $\sigma_j^2 = 1$ et $\tau^2 = 4$. Comparativement à la méthode II, la méthode I produit une amélioration dans la plupart des cas simulés ; le gain maximal est de 30 %, tandis que la seule perte observée est de 9 % pour la combinaison $\sigma_j^2 = 1$ et $\tau^2 = 4$ pour $n = 36$ et $m = 9$. De même, par rapport à la méthode III, le gain maximal donné par la méthode I est de 77 % et la seule perte est de 11 %, pour les mêmes spécifications de paramètres et de tailles d'échantillon.

Comparaisons des PC : Nous avons obtenu les intervalles de confiance au seuil de confiance de 95 %. Les méthodes I et III ne révèlent aucune sous-couverture, ce qui n'est pas étonnant étant donné la construction optimale de leurs intervalles de confiance. La méthode I produit le taux nominal de couverture plus fréquemment que n'importe quelle autre méthode. La méthode II présente une certaine sous-couverture, le taux pouvant être aussi faible que 82 %. *Comparaisons des LMIC* : La méthode I produit en général des intervalles de confiance considérablement plus courts

Tableau 1 Résultats des simulations pour les paramètres du modèle, *a* (panneau supérieur gauche), *b* (panneau supérieur droit), β (panneau inférieur gauche) et τ^2 (panneau inférieur droit). Ici, E.-T. représente l'écart-type sur 200 répliques. Nous avons pris $\beta = 10$ et $\tau^2 = 0,25$, 1 et 4

$n = 36, m = 9$				$n = 180, m = 18$					
τ^2	Moyenne	E.-T.		Moyenne	E.-T.				
0,25	1,0959	0,1540	α	0,25	0,3992	0,0983	b		
	1,0937	0,1555		1	0,4030	0,1012			
4	1,0996	0,1577	β	4	0,3999	0,1017	τ^2		
0,25	10,0071	0,3618		0,25	0,2558	0,0605			
	10,0142	0,3311		1	0,9418	0,3333			
4	10,0282	0,4639		4	3,5592	1,3316			
$n = 36, m = 9$				$n = 180, m = 18$					
τ^2	Moyenne	E.-T.		Moyenne	E.-T.				
0,25	9,9951	0,1853	α	0,25	0,3992	0,0983	b		
1	9,9970	0,1743		1	0,4030	0,1012			
4	10,0048	0,2254	β	4	0,3999	0,1017	τ^2		
0,25	10,0071	0,3618		0,25	0,2558	0,0605			
	10,0142	0,3311		1	0,9418	0,3333			
4	10,0282	0,4639		4	3,5592	1,3316			

Tableau 2 Résultats des simulations pour la prédiction quand $\tau^2 = 0,25$. Ici, EQM, LMIc et PC représentent l'erreur quadratique moyenne, la longueur moyenne de l'intervalle de confiance et la probabilité de couverture de l'intervalle de confiance, respectivement

$n = 36, m = 9$				$n = 180, m = 18$			
σ^2	I	II	Méthode	I	II	Méthode	
1	0,0048	0,0198	III	-0,0051	-0,0086	III	IV
	-0,0033	-0,0061		-0,0130	-0,0109		
4	0,0126	0,0370	II	-0,0045	-0,0045	I	I
16	0,0381	0,5430		0,3000	0,5805		
	0,3715	0,5240	I	0,2850	0,4856	II	II
1	2,1393	2,5485		1,6006	3,6466		
4	2,2632	3,9574	IV	2,1524	5,2472	IV	IV
16	2,3221	4,5619		2,3308	6,5273		
	9,3335	11,1363	III	2,1046	5,5644	III	III
1	0,9771	0,9708		0,9564	0,9710		
4	0,9660	0,9933	II	0,9529	0,9610	II	II
16	0,9710	0,9829		0,9555	0,9660		
	0,9468	0,9710	I	0,9600	0,9998	I	I
1	0,9468	0,9771		0,9851	0,9998		
4	0,9365	0,9660	IV	0,9967	0,9998	IV	IV
16	0,9660	0,9933		0,9975	0,9998		
	0,9812	0,9808	III	0,9918	0,9998	III	III
1	0,9762	0,9775		0,9627	0,9680		
4	0,9490	0,9912	II	0,9680	0,9918	II	II
16	0,9503	0,9812		0,9680	0,9918		
	0,9640	0,9762	I	0,9650	0,9786	I	I
1	0,9704	0,9762		0,9650	0,9786		
4	0,9321	0,9733	IV	0,9323	0,9660	IV	IV
16	4,4082	7,4286		4,5543	7,8937		
	3,1822	4,4938	III	3,6763	7,8937	III	III
1	3,4550	3,2117		2,8676	7,8937		
4	1,0482	2,3109	II	1,1878	7,8937	II	II
16	1,3100	2,1059		1,1024	7,8937		
	0,8566	1,5396	I	0,9415	7,8937	I	I
1	0,5645	0,7238		0,5673	7,8937		
4	0,3715	0,5240	IV	0,3748	7,8937	IV	IV
16	0,3281	0,5430		0,3748	7,8937		
	0,3066	0,3890	III	0,2922	7,8937	III	III
1	0,3066	0,3890		0,2922	7,8937		
4	0,0126	0,0370	II	-0,0061	7,8937	II	II
16	0,0126	0,0370		-0,0061	7,8937		
	-0,0033	-0,0061	I	-0,0116	7,8937	I	I
1	0,0048	0,0198		-0,0112	7,8937		

Tableau 3 Résultats des simulations pour la prédiction quand $\tau^2 = 1$. Ici, EQM, LMIc et PC représentent l'erreur quadratique moyenne, la longueur moyenne de l'intervalle de confiance et la probabilité de couverture de l'intervalle de confiance, respectivement

$n = 36, m = 9$				$n = 180, m = 18$			
σ^2	I	II	Méthode	I	II	Méthode	
16	0,0152	0,0205	III	-0,0051	-0,0085	III	IV
	-0,0167	-0,0164		-0,0121	-0,0133		
4	0,0566	1,1100	II	0,5288	0,5673	II	II
1	0,6545	0,7230		0,5430	0,5673		
16	1,0482	2,1059	IV	0,9766	0,8770	IV	IV
	1,3100	2,3156		1,0000	1,1024		
1	3,4550	4,4938	I	3,1088	3,6763	I	I
4	4,0321	6,8984		3,7847	4,5543		
16	7,4286	9,3555	III	4,1187	6,6785	III	III
	11,1555	11,1555		5,1590	7,8937		
1	0,9762	0,9760	II	0,9660	0,9786	II	II
4	0,9812	0,9808		0,9680	0,9918		
16	0,9533	0,9490	I	0,9680	0,9974	I	I
	0,9560	0,9560		0,9680	0,9974		

est suffisant de considérer les deux cas suivants : i) $\hat{\sigma}_i^2 \geq \hat{\sigma}_i^2(\infty)$, où il s'ensuit que $(n_i + 2a + 2)\hat{\sigma}_i^2 = (n_i + 2a + 1)\hat{\sigma}_i^2 \geq \hat{\sigma}_i^2 \geq (n_i + 2/b + 2)(n_i - 1)S_i^2$, et ii) $\hat{\sigma}_i^2 \geq 1$, où il s'ensuit que $(n_i + 2a + 2)\hat{\sigma}_i^2 \geq (n_i + 2/b + 2)(n_i - 1)S_i^2$, et iii) $\hat{\sigma}_i^2 \geq \hat{\sigma}_i^2(0)$, où il s'ensuit que $(n_i + 2a + 2)\hat{\sigma}_i^2 = (X_i - Z_i^T \beta)^2 + (n_i - 1)S_i^2 \geq 2/b \geq (n_i - 1)S_i^2$. Donc, dans les cas (i) ainsi que (iii),

$$(n_i + 2a + 2)\hat{\sigma}_i^2 \geq (n_i - 1)S_i^2. \quad (19)$$

Puisque $\theta_i - \mu_i \sim N(0, \sigma_i^2 \tau_i^2 / (\sigma_i^2 + \tau_i^2))$ et $(n_i - 1)S_i^2 / \sigma_i^2 \sim \chi_{n_i-1}^2$, l'intervalle de confiance

$$D_i = \left\{ \theta_i : n_i \frac{S_i}{|\theta_i - \mu_i|} \leq t_{\alpha/2} \right\} \quad (20)$$

a une probabilité de couverture de $1 - \alpha$. Donc, si n_0 et μ_i sont remplacés par \hat{n}_0 et $\hat{\mu}_i$, il faut s'attendre à ce que l'intervalle de confiance résultant D_i , disons, ait une probabilité de couverture d'environ $1 - \alpha$. De (19), nous obtenons

$$P\{\hat{C}_i(\mathbf{B})\} \geq P(\hat{D}_i) \approx 1 - \alpha, \quad (21)$$

ce qui établit une borne inférieure approximative de $1 - \alpha$ pour le seuil de confiance de $\hat{C}_i(\mathbf{B})$.

Dans (21), \mathbf{B} était supposé fixe et connu. Quand \mathbf{B} est inconnu, nous le remplaçons par l'estimation de son maximum de vraisemblance marginale $\hat{\mathbf{B}}$. Puisque l'expression (21) est vérifiée quelle que soit la valeur réelle de \mathbf{B} , la substitution de $\hat{\mathbf{B}}$ à \mathbf{B} dans (21) comportera une erreur d'ordre $O(1/\sqrt{N})$, où $N = \sum_{i=1}^n n_i$. Comparativement à chaque n_i , pris individuellement, ce groupement des n_i devrait réduire l'erreur de manière significative, de manière que $\hat{C}_i(\mathbf{B})$ soit suffisamment proche de $C_i(\mathbf{B})$ pour satisfaire la borne inférieure de $1 - \alpha$ dans (21).

5. Une étude par simulation

5.1 Conditions de simulation

Nous considérons les conditions de simulation dans lesquelles nous utilisons un sous-ensemble de configurations d'échantillon emprunté à Wang et Fuller (2003). Chaque échantillon employé dans l'étude par simulation a été obtenu en suivant les étapes que voici. Premièrement, générer des observations en utilisant le modèle

$$X_{ij} = \beta + n_i + e_{ij},$$

où $n_i \sim N(0, \tau_i^2)$ et $e_{ij} \sim N(0, n_i \sigma_i^2)$, indépendamment pour $j = 1, \dots, n_i$ et $i = 1, \dots, n$. Alors, le modèle à effets aléatoires pour la moyenne de petit domaine, X_i , est

$X_i = \beta + n_i + e_i$, indépendamment pour $i = 1, \dots, n$, où $X_i \equiv \bar{X}_i \equiv n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$ et $e_i \equiv \bar{e}_i \equiv n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$. Donc, $X_i \sim N(\theta_i, \sigma_i^2)$, où $\theta_i = \beta + n_i$, $\theta_i \sim N(\beta, \tau_i^2)$ et $e_i \sim N(0, \sigma_i^2)$. Nous avons estimé σ_i^2 en nous servant de l'estimateur sans biais

$$S_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,$$

et il s'ensuit que $(n_i - 1)S_i^2 / \sigma_i^2 \sim \chi_{n_i-1}^2$, indépendamment pour $i = 1, 2, \dots, n$. Notons que le plan de simulation ne tenait pas compte de la modélisation des variances d'échantillonnage au deuxième niveau dans (2). Par conséquent, notre résultat indiquera une robustesse à l'erreur de spécification du modèle de variance.

Les étapes susmentionnées ont produit les données (X_i, S_i^2) , $i = 1, \dots, n$. Pour simplifier la simulation, nous ne choisissons aucune covariable \mathbf{Z}_i . À l'instar de Wang et Fuller (2003), nous donnons la valeur m à tous les n_i afin de faciliter la programmation. Cependant, nous choisissons quand même que les variances d'échantillonnage réelles soient inégales : la valeur d'un tiers des σ_i^2 est fixée à 1, celle d'un deuxième tiers est fixée à 4 et celle du dernier tiers est fixée à 16. Nous prenons $\beta = 10$ et trois valeurs différentes de $\tau_i^2 = 0.25$, 1 et 4. Nous avons choisi ces valeurs des paramètres en nous inspirant de Qiu et Hwang (2007). Pour chaque valeur de τ_i^2 , nous avons généré 200 échantillons pour les deux combinaisons $(m, n) = (9, 36)$ et $(18, 180)$.

Dans l'étude par simulation, nous comparons la méthode (EQM), la probabilité de couverture (PC) des intervalles de confiance et la longueur moyenne des intervalles de confiance (LMC). Le tableau 1 donne les estimations des paramètres pour a , b , β et τ_i^2 . Les résultats numériques indiquent que les estimations du maximum de vraisemblance des paramètres du modèle ont de bonnes propriétés ; les valeurs estimées de β et τ_i^2 sont proches des valeurs réelles, ce qui témoigne de bonnes propriétés de robustesse à l'erreur de spécification de la distribution au deuxième niveau de (2). L'obtention d'estimations statistiquement significatives pour a ainsi que b indique que les variances d'échantillonnage « rétrogrades » sont intégrées dans la méthode proposée. Les tableaux 2, 3 et 4 donnent les moyennes des résultats numériques calculées sur les domaines qui, dans chaque groupe, ont les mêmes variances mais qui, dans chaque groupe, ont les mêmes variances fondées sur 200 répétitions.

Remarque 1. Hwang et coll. (2009) ont choisi k en prenant (15) égale à l'intervalle t fondé sur X_i^j seulement pour les paramètres de petit domaine θ_i . Notons que X_i^j est l'estimateur direct d'après les données d'enquête. Par conséquent, ce choix de k n'exerce aucun contrôle direct sur la probabilité de couverture de l'intervalle construit sous estimation par rétrocissement. Par ailleurs, notre choix proposé de k a été établi de manière à maintenir la couverture nominale sous, précisément, l'estimation par rétrocissement.

Remarque 2. Notons qu'en l'absence de toute hypothèse de modélisation hiérarchique, S_i^j et X_i^j sont indépendants car S_i^j et X_i^j sont, respectivement, auxiliaire et la statistique exhaustive complète pour θ_i . Cependant, sous les modèles (1) et (2), la distribution conditionnelle de σ_i^2 et θ_i fait intervenir à la fois X_i^j et S_i^j , ce que l'on peut constater en examinant (5) et (9).

Remarque 3. Dans Hwang et coll. (2009), l'estimateur à rétrocisseur de σ_i^2 est fondé uniquement sur l'information au sujet de S_i^j et non au sujet de X_i^j ainsi que S_i^j . L'estimateur bayésien de σ_i^2 est introduit par insertion dans l'expression de l'estimateur bayésien des paramètres de petit domaine. Donc, l'estimateur sur petits domaines de Hwang et coll. s'écrit sous la forme $E(\theta_i | X_i^j, \hat{\sigma}_{i,B}^2)$ dans (14) où $\hat{\sigma}_{i,B}^2$ est l'estimateur bayésien de σ_i^2 . En raison de l'équation (9), l'estimateur à rétrocisseur de σ_i^2 dépend de $(X_i^j - Z_i^j \beta)^T$ en plus de S_i^j contrairement à l'estimateur de Hwang et coll. (2009). Nous pensons que cela pourrait être l'explication de la meilleure performance de notre méthode comparativement à celle de Hwang et coll. (2009).

Remarque 4. Comme nous l'avons mentionné plus haut, le nombre de degrés de liberté associés à la distribution χ^2 pour la variance d'échantillonnage ne doit pas être simplement $n_i - 1$, n_i étant la taille de l'échantillon pour le i^e domaine. Il n'existe aucun résultat théorique fiable pour déterminer le nombre de degrés de liberté quand le plan de sondage est complexe. Wang et Fuller (2003) ont approximé la distribution χ^2 par une distribution normale fondée sur l'approximation de Wilson-Hilferty. Si l'on connaît le plan de sondage exact, les lignes directrices basées sur la simulation de Mäples et coll. (2009) pourraient être utiles. Pour produire des estimations au niveau du comté en se servant des données de l'American Community Survey, Mäples et coll. (2009) ont suggéré d'estimer le nombre de degrés de liberté par $0,36 \times \sqrt{n_i}$.

4. Justification théorique

À la présente section, nous donnons la justification théorique du choix de k suivant l'équation (10). Comme

$$\mu_i = w_i X_i^j + (1 - w_i) Z_i^j \beta, \quad (17)$$

$$v_i = \left(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_i^2} \right)^{-1} = \sigma_i^2 \left(1 + \frac{1}{\sigma_i^2} \right)^{-1},$$

dans Hwang et coll. (2009), la distribution conditionnelle de θ_i sachant X_i^j et S_i^j peut être approximée par $\pi(\theta_i | X_i^j, S_i^j, \mathbf{B}) \approx \pi(\theta_i | X_i^j, S_i^j, \mathbf{B}, \hat{\sigma}_i^2)$, où $\hat{\sigma}_i^2$ est défini comme dans (11). De la même façon, nous approximations $E(\sigma_i^{-1} | X_i^j, S_i^j, \mathbf{B})$ par $E(\sigma_i^{-1} | X_i^j, S_i^j, \mathbf{B}) \approx \hat{\sigma}_i^{-1}$. Sur la base de ces approximations, nous avons $C_i(\mathbf{B}) \approx C_i(\mathbf{B})$ où $C_i(\mathbf{B})$ est l'intervalle de confiance de θ_i donné par $C_i(\mathbf{B}) = \{\theta_i : \pi(\theta_i | X_i^j, S_i^j, \mathbf{B}, \hat{\sigma}_i^2) \geq k \hat{\sigma}_i^{-1}\}$. De (1) il découle que la densité de probabilité conditionnelle $\pi(\theta_i | X_i^j, S_i^j, \mathbf{B}, \sigma_i^2)$ est normale de moyenne μ_i et de variance v_i , où μ_i et v_i sont donnés par les expressions

$$w_i = \frac{1 / \sigma_i^2 + 1 / \tau^2}{1 / \sigma_i^2 + 1 / \tau^2 + 1},$$

Maintenant, en choisissant

$$k = n_0 \phi \left(t_{\alpha/2} \sqrt{\frac{n_i - 2a + 2}{n_i - 1}} \right)$$

comme nous l'avons mentionné, l'intervalle de confiance $C_i(\mathbf{B})$ devient

$$C_i(\mathbf{B}) = \left\{ \theta_i : t_{\alpha/2} \left| \frac{\hat{\theta}_i}{\theta_i - \hat{\mu}_i} \right| \leq t_{\alpha/2} \sqrt{\frac{n_i - 1}{n_i - 2a + 2}} \right\}, \quad (18)$$

où $\hat{\mu}_i$ est l'expression de μ_i dans (17) avec remplacement de $\hat{\sigma}_i^2$ par $\hat{\sigma}_i^2(\mathbf{B})$. Considérons maintenant le comportement de $\hat{\sigma}_i^2 \equiv \hat{\sigma}_i^2(\mathbf{B})$ quand τ^2 varie entre 0 et ∞ . Quand $\tau^2 \rightarrow \infty$, $\hat{\sigma}_i^2$ converge vers

$$\hat{\sigma}_i^2(\infty) \equiv \hat{\sigma}_i^2(a, b, \beta, \infty) = \frac{\frac{2}{n_i - 1} + a + 1}{\frac{2}{n_i - 1} + 1} = \frac{2}{n_i - 1} S_i^j + \frac{2}{b}.$$

De même, quand $\tau^2 \rightarrow 0$, $\hat{\sigma}_i^2$ converge vers

$$\hat{\sigma}_i^2(0) \equiv \hat{\sigma}_i^2(a, b, \beta, 0) = \frac{(X_i^j - Z_i^j \beta)^T + (n_i - 1) S_i^j + \frac{2}{b}}{n_i + 2a + 2}.$$

Pour toute valeur intermédiaire de τ^2 , nous avons $\min\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\} \leq \hat{\sigma}_i^2 \leq \max\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\}$. Donc, il

$$(12) \quad \begin{cases} X_i | \theta_i, \sigma_i^2 \sim \text{Normale}(\theta_i, \sigma_i^2) \\ \theta_i \sim \text{Normale}(\mathbf{Z}_i^T \boldsymbol{\beta}, \tau_i^2) \end{cases}$$

$$(13) \quad \begin{cases} \log S_i^2 = \log(\sigma_i^2) + \delta_i, \quad \delta_i \sim N(m_i, \sigma_{ch,i}^2) \\ \log(\sigma_i^2) \sim N(\mu_i, \tau_i^2) \end{cases}$$

indépendamment pour $i = 1, 2, \dots, n$. Notons que le modèle de la moyenne dans (12) est identique à celui figurant dans (1). Les quantités τ_i^2 , m_i et $\sigma_{ch,i}^2$ sont supposées connues et sont données par $m_i = E[\log(\chi_{n-1}^2/(n-1))] = \log 2$ et $\sigma_{ch,i}^2 = \text{Var}[\log(\chi_{n-1}^2/(n-1))] = 2/(n-1)$. Donc, la taille d'échantillon n_i détermine la forme de la distribution χ^2 par la voie du paramètre de nombre de degrés de liberté mais surtout, comme nous l'avons mentionné plus haut, les tailles d'échantillon différentes expliquent différents degrés de réajustement du paramètre de variance réelle correspondant dans leur approche d'estimation, les paramètres μ_i et τ_i^2 inconnus du modèle sont estimés selon une méthode fondée sur le moment dans un cadre bayésien empirique dominant μ_i et τ_i^2 respectivement. Notons que, dans Hwang et coll. (2009), des estimations sont obtenues en se basant sur le modèle hiérarchique pour σ_i^2 dans (13) *entièrement*, sans se préoccuper de la modélisation (1) de la moyenne. Nous renvoyons le lecteur à la section 5 de leur article pour des renseignements détaillés sur l'estimation des hyperparamètres. Nous suivons la même procédure en utilisant uniquement (13) pour estimer μ_i et τ_i^2 dans le cas de tailles d'échantillon inégales.

La dérivation de l'estimation bayésienne de σ_i^2 est

$$\hat{\sigma}_{i,B}^2 = \exp \left[E[\ln(\sigma_i^2)] \ln(S_i^2) \right]$$

$$= \left\{ \frac{S_i^2}{m_i} \exp\{\mu_i(1 - \tau_i^2)\} M_{\tau_i^2} \right\}$$

où $M_{\tau_i^2} = \tau_i^2 / (\tau_i^2 + \sigma_i^2)$ et avec insertion des estimations pour remplacer les quantités inconnues. La distribution conditionnelle de θ_i sachant (X_i, S_i^2) , qui est donnée par

$$\pi(\theta_i | X_i, S_i^2) = \int_0^\infty \pi(\theta_i | X_i, S_i^2, \sigma_i^2) \pi(\sigma_i^2 | X_i, S_i^2) d\sigma_i^2,$$

est approximée par $\pi(\theta_i | X_i, S_i^2) \approx \int_0^\infty \pi(\theta_i | X_i, S_i^2, \hat{\sigma}_{i,B}^2) \pi(\sigma_i^2 | X_i, S_i^2) d\sigma_i^2 = \pi(\theta_i | X_i, S_i^2)$. Cela suggère l'estimateur bayésien approximatif des paramètres de petit domaine donné par

$$(14) \quad \theta_i = E(\theta_i | X_i, \hat{\sigma}_{i,B}^2) = \hat{M}_i X_i + (1 - \hat{M}_i) \mathbf{Z}_i^T \boldsymbol{\beta},$$

où $\hat{M}_i = \tau_i^2 / (\tau_i^2 + \hat{\sigma}_{i,B}^2)$. L'intervalle de confiance pour θ_i s'obtient sous la forme

$$(15) \quad C_H^i = \left\{ \theta_i : \left| \frac{\theta_i - \hat{\theta}_i}{\hat{\sigma}_{i,B}^2} \right| < -2 \ln \{k\sqrt{2\pi}\} - \ln(\hat{M}_i) \right\},$$

À la section 3 de Hwang et coll. (2009), pages 269 à 271, l'intervalle C_H^i est associé avec l'intervalle t à 100(1- α)% $[|\theta_i - X_i| < t S_i]$ pour obtenir l'expression de k comme $k \equiv k_i = \exp\{-t^2/2\} \exp\{m_i/2\} / (\sqrt{2\pi})$. *Méthode IV* : Cette méthode comprend un cas particulier du modèle de Fay-Herriot donné en (1), mais avec l'estimation des paramètres du modèle empruntée à Qiu et Hwang (2007). Ces derniers ont considéré le modèle

$$(16) \quad \begin{cases} X_i | \theta_i, \sigma_i^2 \sim \text{Normale}(\theta_i, \sigma_i^2) \\ \theta_i \sim \text{Normale}(0, \tau_i^2) \end{cases}$$

indépendamment pour $i = 1, 2, \dots, n$, pour analyser des données micro vectorielles expérimentales. Dans le cas où les paramètres du modèle étaient connus, ils ont proposé l'estimateur ponctuel $\hat{\theta}_i = \hat{M}_i X_i$ où $\hat{M}_i = (1 - (n - 2)\sigma_i^2 / |X_i|^2)^+$, où $a^+ = \max(a, 0)$ pour tout nombre a et $|X_i| = |(\sum_{j=1}^n X_{ij}^2)^{1/2}|$. L'intervalle de confiance pour θ_i est $\hat{\theta}_i \pm v_i^*(\hat{M}_i)$, où $v_i^*(\hat{M}_i) = \sigma_i^2 \hat{M}_i (q_1 - \ln(\hat{M}_i))$ avec q_1 désignant la valeur critique de la variable normale standard pour le niveau de confiance souhaité et $v_i(0) = 0$. Ici, en vue de procéder à la comparaison avec notre méthode, nous modifions le premier niveau du modèle hiérarchique dans (16) comme il suit :

$$X_i = \mathbf{Z}_i^T \boldsymbol{\beta} + v_i + e_i$$

où $v_i \sim \text{Normale}(0, \tau_i^2)$ et $e_i \sim \text{Normale}(0, S_i^2)$ indépendamment pour $i = 1, 2, \dots, n$, et S_i^2 est traité comme étant connu. Comme Qiu et Hwang (2007), nous estimons τ_i^2 par

$$\hat{\tau}_i^2 = \frac{1}{d} \left[\sum_{i=1}^n \hat{\sigma}_i^2 - \sum_{i=1}^n \sum_{j=1}^n \left(\mathbf{Z}_i^T \mathbf{Z}_j^T \right)^{-1} \left(\sum_{j=1}^n \mathbf{Z}_i^T \mathbf{Z}_j^T \right) \right]$$

et $\hat{\tau}_i^2 = \max(\hat{\tau}_i^2, 1/n)$, où $\hat{\sigma}_i^2 = X_i - \mathbf{Z}_i^T \boldsymbol{\beta}$ et $\hat{\beta} = (\sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T)^{-1} \sum_{i=1}^n \mathbf{Z}_i Y_i$. Puis nous définissons $\hat{M}_{0i} = \hat{\tau}_i^2 / (\hat{\tau}_i^2 + S_i^2)$ et $\hat{M}_i = \max(\hat{M}_{0i}, M_i)$, où, dans la dernière expression, \hat{M}_{0i} est tronqué par $M_i = 1 - \hat{\sigma}_i^2 / (n - 2)$, et $\hat{\sigma}_i^2$ est le α^e quantile d'une distribution du khi-carré à n_i degrés de liberté. Cet \hat{M}_i est utilisé dans la formule de l'intervalle de confiance de θ_i donnée plus haut. Quand nous avons appliqué cette méthode dans notre étude par simulation et notre analyse des données réelles, nous avons modifié le modèle afin de pouvoir utiliser les tailles d'échantillon inégales et l'information sur les covariables mentionnées plus haut.

ensemble crédible bayésien. Cependant, à l'instar de Casella et Berger (1990, page 470), nous continuerons de donner à $C_i(\mathbf{B})$ le nom d'intervalle de confiance. Dans une perspective bayésienne empirique également, cette terminologie est plus appropriée. Nous montrerons à la section 3.3 comment le paramètre de mise au point k détermine le niveau de confiance de $C_i(\mathbf{B})$.

En supposant pour le moment que k est connu, nous suivons les étapes ci-après pour calculer $C_i(\mathbf{B})$. Les densités conditionnelles de σ_i^2 et θ_i sont données par

$$\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \propto$$

$$(9) \quad \frac{\exp \left[-0.5(X_i - \mathbf{Z}_i^T \mathbf{B})^2 / \left\{ 0.5(n_i - 1)S_i^2 + \frac{1}{b} \right\} \left(\frac{\sigma_i^2}{1} \right) \right]}{(\sigma_i^2)^{(n_i-1)/2+a+1} (\sigma_i^2 + \tau_i^2)^{1/2}}$$

et (5), respectivement, expressions qui, comme nous l'avons mentionné plus haut, n'ont pas de forme analytique. Donc, comme dans le cas de θ_i , nous calculons $E(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$ numériquement en utilisant la méthode Monte Carlo par approximation de la valeur prévue de la moyenne $1/N \sum_{N=1}^K 1/\sigma_{i,p}^2$ où $\sigma_{i,p}^2, p = 1, 2, \dots, R$ sont les R échantillons tirés de la densité conditionnelle $\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$. La procédure d'acceptation-rejet est utilisée pour tirer des nombres aléatoires de $\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$ avec une loi instrumentale donnée par la loi Gamma inverse

$$\exp \left[- \left\{ 0.5(n_i - 1)S_i^2 + \frac{1}{b} \right\} \left(\frac{\sigma_i^2}{1} \right) \right] \frac{(\sigma_i^2)^{(n_i-1)/2+a+1}}{}$$

et la probabilité d'acceptation

$$\exp \left\{ -0.5(X_i - \mathbf{Z}_i^T \mathbf{B})^2 / \left\{ 0.5(n_i - 1)S_i^2 + \frac{1}{b} \right\} \left(\frac{\sigma_i^2}{1} \right) \right\} \times \exp(0.5) \times |X_i - \mathbf{Z}_i^T \mathbf{B}|.$$

L'étape suivante consiste à déterminer les valeurs des bornes de $C_i(\mathbf{B})$ en trouvant deux valeurs de θ_i qui satisfont l'équation $kE(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) - \pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) = 0$. Il faut pour cela que la constante de normalisation donnée en (5)

$$D_i = \int_{-\infty}^{\infty} \exp \{ -0.5(\theta_i - \mathbf{Z}_i^T \mathbf{B})^2 / \tau_i^2 \} \psi_i^{(n_i/2+a)} d\theta_i$$

soit évaluée numériquement. Nous le faisons en procédant à l'intégration de Gauss-Hermite avec 20 nœuds.

3.3 Choix de k

Nous choisissons pour le paramètre de mise au point k dans (8) l'expression

$$(10) \quad k = k(\mathbf{B}) = n_{i,0} \phi \left(t_{\alpha/2} \sqrt{\frac{n_i + 2a + 2}{n_i - 1}} \right)$$

où ϕ est la distribution normale standard, $t_{\alpha/2}$ est le $(1 - \alpha/2)^{\text{e}}$ centile de la distribution t avec $n_i - 1$ de-
grés de liberté, et $n_{i,0} = \sqrt{1 + \sigma_i^2 / \tau_i^2}$. Puisque $n_{i,0}$ fait inter-
venir σ_i^2 qui est inconnue, une version estimée $\hat{n}_{i,0}$ s'ob-
tient en introduisant l'estimation du maximum a posteriori

$$(11) \quad \hat{\sigma}_i^2 = \hat{\sigma}_i^2(\mathbf{B}) = \arg \max_{\sigma_i^2} \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \Big|_{\mathbf{B}=\mathbf{B}}$$

à la place de σ_i^2 . En outre, nous remplaçons \mathbf{B} par $\hat{\mathbf{B}}$ dans (11). Nous démontrons que la probabilité de couverture de $C_i(\mathbf{B})$ avec ce choix de k s'approche de $1 - \alpha$. Les justi-
fications théoriques sont présentées à la section 4.

3.4 Autres méthodes apparentées aux fins de comparaison

Nous donnons à notre méthode le nom de méthode L. Nous décrivons brièvement ci-dessous trois autres méthodes auxquelles nous la comparerons.

Méthode II : Wang et Fuller (2003) ont considéré le modèle d'estimation sur petits domaines de Fay-Herriot donné par (1). Leur principale contribution est la construction de la formule d'estimation de l'erreur quadratique moyenne pour les estimateurs sur petits domaines avec variances d'échantillonnage estimées. Ce faisant, ils ont construit deux formules désignées par EQM₁ et EQM₂. Pour nos comparaisons, nous utilisons EQM₁, qui a été dérivée en suivant l'approche de correction du biais de Prasad et Rao (1990). La différence fondamentale par rapport à notre approche est qu'ils n'ont pas lissé les variances d'échantillonnage, et n'ont tenu compte de l'incertitude que dans l'inférence au sujet des paramètres de petit domaine. La méthode d'estimation des paramètres du modèle est fondée sur les moments pour tous les paramètres du modèle, diffère également de la nôtre.

Méthode III : Hwang et coll. (2009) ont considéré les modèles log-normal et Gamma inverse pour σ_i^2 dans (2) pour l'analyse des données micro vectorielles. Leur étude par simulation a montré que les propriétés des intervalles de confiance des estimateurs sur petits domaines étaient meilleures sous modèle log-normal que sous le modèle Gamma inverse. Nous avons donc modifié leur modèle log-normal afin d'ajouter des covariables et des tailles d'échan-
tillon n_i inégales comme il suit :

recourons à une méthode de Monte Carlo pour évaluer l'expression. Supposons que R échantillons iid de θ_i soient disponibles, disons $\theta_{i,1}, \theta_{i,2}, \dots, \theta_{i,R}$. Alors, chaque expression de la forme $E\{h(\theta_i)\}$ peut être approximée par la moyenne Monte Carlo

$$(6) \quad E\{h(\theta_i)\} \approx \frac{1}{R} \sum_{r=1}^R h(\theta_{i,r}).$$

Nous posons que $(B^{(t)}) = (a^{(t)}, b^{(t)}, \beta^{(t)}, \tau^{(t)})^T$ et procédons à la $(t+1)^{\text{e}}$ étape. Cette procédure de maximisation est répétée jusqu'à la convergence de l'estimation $(B^{(t)})$. L'EMV de B est $B^{(t)}$ une fois que la convergence est établie.

3.2 Estimation ponctuelle et intervalle de confiance de θ_i

Selon la technique classique, nous posons que l'estimateur sur petits domaines de θ_i est

$$(7) \quad \hat{\theta}_i = E(\theta_i | X_i, S_i^2, Z_i, B) \Big|_{B=B^*}$$

L'espérance de θ_i par rapport à la densité conditionnelle $\pi(\theta_i | X_i, S_i^2, Z_i, B)$ où l'estimation du maximum de vraisemblance B est « insérée » pour remplacer B . L'estimation $\hat{\theta}_i$ est calculée numériquement en utilisant la procédure Monte Carlo (6) décrite à la section précédente. Dans la suite, le paramètre B inconnu sera remplacé par B^* dans toutes les quantités dans lesquelles il intervient, même si nous continuons d'utiliser la notation B pour simplifier.

En outre, nous élaborons un intervalle de confiance pour θ_i fondé sur une théorie de la décision. Comme l'ont fait Joshi (1969), Casella et Hwang (1991), Hwang et coll. (2009), considérons la fonction de perte associée à l'intervalle de confiance C donnée par $(k/\sigma)L(C) - I_C(\theta)$, où k est un paramètre de mise au point. Indépendant des paramètres du modèle, $L(C)$ est la longueur de C et $I_C(\theta)$ est la fonction indicatrice prenant la valeur 1 ou 0 selon que $\theta \in C$ ou non. Notons que cette fonction de perte tient compte à la fois de la probabilité de couverture et de la longueur de l'intervalle; la quantité positive (k/σ) sert de poids relatif de la longueur comparativement à la probabilité de couverture de C est $(-\infty, \infty)$ avec une sorte que la valeur optimale de C est $(-\infty, \infty)$ avec une probabilité de couverture de 1. Par ailleurs, pour $k = \infty$, la probabilité de couverture est égale à 0, de sorte que la valeur optimale de C est un ensemble de points. Pour obtenir l'intervalle de confiance de Bayes de θ_i , il faut minimiser la fonction de risque $E\{[(k/\sigma)L(C) - I_C(\theta)] | X_i, S_i^2, Z_i, B\}$. Le choix optimal de C est donné par

$$C_i(B) = \{\theta_i : kE(\sigma_i^{-1} | X_i, S_i^2, Z_i, B) > \pi(\theta_i | X_i, S_i^2, Z_i, B)\}. \quad (8)$$

Puisque que $C_i(B)$ s'obtient en minimisant le risque a posteriori, on pourrait vouloir l'interpréter comme un

études que nous avons effectuées. Le maximiseur de $\bar{Q}(B|B^{(t-1)})$ à la t^{e} étape peut être décrit explicitement. Les solutions pour β et τ^2 sont disponibles sous les formes analytiques suivantes

$$\beta^{(t)} = \left(\sum_{i=1}^n Z_i' Z_i' E(\theta_i) \right) \left(\sum_{i=1}^n Z_i' Z_i' E(\theta_i) \right)^{-1} \quad \text{et} \quad (\tau^2)^{(t)} = \frac{1}{n} \sum_{i=1}^n E(\theta_i) - Z_i' \beta^{(t)},$$

respectivement. En outre, $a^{(t)}$ et $b^{(t)}$ s'obtiennent en résolvant $S_a^{(t)} = \partial \bar{Q}(B|B^{(t-1)}) / \partial a = 0$ et $S_b^{(t)} = \partial \bar{Q}(B|B^{(t-1)}) / \partial b = 0$ par la méthode de Newton-Raphson où

$$S_a^{(t)} = \frac{\partial}{\partial a} \log \{ \Gamma(n_i/2 + a) \}$$

$$- n \left\{ \frac{\partial}{\partial a} \log \{ \Gamma(a) \} - n \log(b) \right\} - \sum_{i=1}^n E \{ \log(\psi_i) \}$$

doit être déterminé prudemment (par exemple, Maples et coll. 2009). Surtout, le rôle de taille d'échantillon dans l'estimation par rétrécissement de σ_i^2 est le suivant : l'estimation de σ_i^2 se rapproche davantage de la moyenne globale (ab) pour de faibles valeurs de n_i que pour des valeurs élevées. Donc, pour les variances, les tailles d'échantillon jouent le même rôle que la précision dans l'estimation par rétrécissement des moyennes de petit domaine. Nous notons que You et Chapman (2006) ont également envisagé un deuxième niveau de modélisation de la variance d'échantillonnage. Cependant, les hyperparamètres reliés à la loi a priori de σ_i^2 ne sont pas dictés par les données mais plutôt choisis de façon telle que la loi a priori soit vague. Donc, leur modèle peut être considéré comme la version bayésienne de modèle examiné dans Rives et Vandal (2003) et dans Wang et Fuller (2003). Le deuxième niveau de modélisation de σ_i^2 dans (2) peut être étendu encore davantage à $\sigma_i^2 \sim \text{Gamma}(b, \exp(\mathbf{Z}_i' \beta_2) / b)$ de sorte que $E(\sigma_i^2) = \exp(\mathbf{Z}_i' \beta_2)$ pour un autre jeu de p coefficients de régression β_2 afin d'inclure l'information sur les covariables dans la modélisation de la variance.

Bien que notre modèle soit motivé par Hwang, Qiu et Zhao (2009), nous tenons à mentionner que Hwang et coll. (2009) ont considéré les moyennes et variances par rétrécissement dans le contexte de données micro vectorielles où ils ont préconisé une solution importante consistant à insérer un estimateur à rétrécissement de la variance dans l'estimateur de la moyenne. L'estimateur par rétrécissement de la variance dans Hwang et coll. (2009) est une fonction de S_i^2 seulement et non de X_i ainsi que S_i^2 ; voir les remarques 2 et 3 à la section 2. Donc, l'inférence de la moyenne ne tient pas compte de toute l'incertitude dans l'estimation de la variance. En outre, leur modèle ne contient aucune information sur les covariables. L'étude par simulation décrite plus loin indique que notre méthode d'estimation donne de meilleurs résultats que celle de Hwang et coll. (2009).

Dans la formulation du modèle susmentionné, l'inférence pour le paramètre θ_i représentant la moyenne de petit domaine peut être faite en se basant sur la distribution conditionnelle de θ_i sachant toutes les données $\{(X_i, S_i^2, \mathbf{Z}_i') : i = 1, \dots, n\}$. Sous notre modèle, la distribution conditionnelle de θ_i est une distribution non standard qui ne possède pas de forme analytique et requiert donc des méthodes numériques, telles que la méthode de Monte Carlo et l'algorithme EM, pour l'inférence. Des renseignements détaillés sont fournis à la section suivante.

3. Méthodologie d'inférence

3.1 Estimation des paramètres inconnus au moyen de l'algorithme EM

En pratique, $\mathbf{B} = (a, b, \beta, \tau^2)'$ est inconnu et doit être estimé d'après les données $\{(X_i, S_i^2, \mathbf{Z}_i') : i = 1, 2, \dots, n\}$.

Nous proposons d'estimer \mathbf{B} par la méthode du maximum de vraisemblance marginale : estimer \mathbf{B} ou $\mathbf{B}^{(l)}$ de $L_{M,l}$ la vraisemblance marginale $L_{M,l}(\mathbf{B}) = \prod_{i=1}^n L_{M,i}$ où $L_{M,i} \propto \frac{\Gamma(n_i/2 + a)}{\Gamma(n_i/2 + a)} \int \exp\left\{-\frac{\tau^2}{2}(\theta_i - \mathbf{Z}_i' \beta)^2\right\} \psi_i^{-(n_i/2 + a)} d\theta_i$ (3) $\psi_i \equiv \left\{0.5(X_i - \theta_i)^2 + 0.5(n_i - 1)S_i^2 + \frac{b}{1}\right\}$ (4) La vraisemblance marginale $L_{M,l}$ contient des intégrales qui ne peuvent pas être évaluées en forme analytique, de sorte que l'on doit recourir à des méthodes numériques pour sa maximisation. L'un de ces algorithmes est la procédure itérative EM (espérance-maximisation) qui est utilisée quand on a affaire à ce genre d'intégrales. L'algorithme EM comprend l'augmentation de la vraisemblance observée $L_M(\mathbf{B})$ présentant des données manquantes : dans notre cas, les variables de l'intégration, $\theta_i, i = 1, 2, \dots, n$, constituent cette information manquante. Sachant $\theta \equiv \{\theta_1, \theta_2, \dots, \theta_n\}$, la log-vraisemblance (ℓ) sous données complètes peut s'écrire

$$\ell(\mathbf{B}, \theta) = \sum_{i=1}^n \left[\log\{\Gamma(n_i/2 + a)\} - \log\{\Gamma(a)\} - a \log(b) - 0.5 \log(\tau^2) - \frac{(\theta_i - \mathbf{Z}_i' \beta)^2}{2\tau^2} - (n_i/2 + a) \log(\psi_i) \right],$$

où l'expression de ψ_i est donnée par l'équation (4). Partant d'une valeur initiale de \mathbf{B} , disons $\mathbf{B}^{(0)}$, l'algorithme EM d'exécute itérativement une maximisation par rapport à \mathbf{B} . À la t^e étape, la fonction d'objectif maximisée est

$$\begin{aligned} \tilde{Q}(\mathbf{B} | \mathbf{B}^{(t-1)}) &= E(\ell | \mathbf{B}^{(t-1)}, \theta) \\ &= \sum_{i=1}^n \left[\log\{\Gamma(n_i/2 + a)\} - \log\{\Gamma(a)\} - a \log(b) - 0.5 \log(\tau^2) - \frac{E(\theta_i - \mathbf{Z}_i' \beta)^2}{2\tau^2} - (n_i/2 + a) E\{\log(\psi_i)\} \right]. \end{aligned}$$

Dans $\tilde{Q}(\mathbf{B} | \mathbf{B}^{(t-1)})$, l'espérance est prise par rapport à la distribution conditionnelle de chaque θ_i sachant les données, $\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i', \mathbf{B}^{(t-1)})$, ce qui est

$$\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i', \mathbf{B}) \propto \exp\{-0.5(\theta_i - \mathbf{Z}_i' \beta)^2 / \tau^2\} \psi_i^{-(n_i/2 + a)}. \quad (5)$$

L'une des difficultés ici est que les espérances ne sont pas disponibles sous une forme analytique. Donc, nous

répétée des paramètres du modèle. Nous produisons des intervalles de confiance pour les moyennes dans une perspective de théorie de la décision. La construction des intervalles de confiance est facile à mettre en œuvre en pratique.

La présentation de la suite de l'article est la suivante. Le modèle hiérarchique proposé pour les moyennes et les variances d'échantillonnage est élaboré à la section 2. L'estimation des paramètres du modèle au moyen de l'algorithme EM est exposée à la section 3. La justification théorique des intervalles de confiance proposés et leurs propriétés de couverture sont présentées à la section 4. Une étude par simulation et un exemple fondé sur des données réelles sont présentés aux sections 5 et 6, respectivement. Enfin, une discussion et certaines conclusions sont présentées à la section 7. Une autre formulation du modèle pour les petits domaines, ainsi que des détails mathématiques sont donnés en annexe.

2. Modèle proposé

Supposons que l'on examine n petits domaines. Pour le i^{e} petit domaine, soit de (X_i, S_i^2) la paire comprenant l'estimation directe et la variance d'échantillonnage, pour $i = 1, 2, \dots, n$. Soit $\mathbf{Z}_i = (Z_{i1}^{(p)}, \dots, Z_{iD}^{(p)})^T$ le vecteur de p covariables disponibles à l'étape de l'estimation pour le i^{e} petit domaine. Nous proposons le modèle hiérarchique suivant :

$$(1) \quad \begin{cases} X_i | \theta_i, \sigma_i^2 \sim \text{Normale}(\theta_i, \sigma_i^2) \\ \theta_i \sim \text{Normale}(\mathbf{Z}_i^T \boldsymbol{\beta}, \tau^2) \end{cases} \quad (1)$$
$$(2) \quad \begin{cases} \sigma_i^2 \sim \chi_{n_i-1}^2 \\ \sigma_i^2 \left| \frac{\sigma_i^2}{(n_i-1)S_i^2} \right. \sim \text{Gamma}(a, b) \end{cases} \quad (2)$$

indépendamment pour $i = 1, 2, \dots, n$. Dans l'élaboration du modèle, n_i est la taille d'un échantillon aléatoire simple (EAS) tiré du i^{e} domaine, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ est le vecteur de dimension $p \times 1$ des coefficients de régression, et $\mathbf{B} = (a, b, \beta_1, \tau^2)^T$ est la série complète de paramètres inconnus dans le modèle. En outre, $\text{Gamma}(a, b)$ est la densité de probabilité Gamma dont les paramètres de forme et d'échelle a et b , respectivement, sont positifs, définis comme étant $f(x) = \{b^a \Gamma(a)\}^{-1} e^{-bx} x^{a-1}$ pour $x > 0$, et 0 autrement. Le terme σ_i^2 inconnu est la variance réelle de X_i et est habituellement estimé par la variance d'échantillon S_i^2 . On suppose généralement que les S_i^2 suivent une loi du khi-carré possédant $(n_i - 1)$ degrés de liberté (en raison de la normalité et de l'EAS), mais nous notons que sous des plans de sondage complexes, le nombre de degrés de liberté

sont résumés élégamment dans un article publié en 2008 par William Bell, du *United States Census Bureau*. Ce dernier a examiné minutieusement les conséquences de ces problèmes dans le contexte de l'estimation de l'EQM des estimations de l'EQM pour les petits domaines fondés sur un modèle. Il a également donné des preuves numériques de l'estimation de l'EQM pour le modèle de Fay-Herriot (donné dans l'équation 1) quand il est supposé que les variances d'échantillonnage sont connues. Les progrès exposés jusqu'à présent dans la littérature traitant des petits domaines peuvent être considérés « grosso modo » comme étant i) le lissage des estimations directes des variances des erreurs d'échantillonnage pour obtenir des estimations des variances plus stables dont le biais est faible et ii) la prise en compte (partielle) de l'incertitude dans les variances d'échantillonnage en éten-

nant le modèle de Fay-Herriot.

Malheureusement, l'effort en vue de bien tenir compte des variances d'échantillonnage dans la modélisation de la moyenne a été faible, voire nul, comparativement au nombre d'études consacrées à la modélisation et à l'inférence des moyennes. Le développement systématique du « réticissement » des moyennes ainsi que des variances fait défaut dans la littérature traitant de l'estimation sur petits domaines. Autrement dit, nous aimerions exploiter la technique de l'« emprunt d'information » à d'autres petits domaines en vue d'« améliorer » les estimations de la variance, tout comme nous le faisons pour « améliorer » les estimations des moyennes de petits domaines. Nous proposons un modèle hiérarchique utilisant à la fois les estimations directes d'après les données d'enquête et les estimations des variances d'échantillonnage pour inférer les paramètres du modèle qui déterminent le système stochastique. Notre objectif méthodologique est d'élaborer l'estimation « par réticissement » double pour les moyennes de la modélisation conjointe moyenne-variance afin que les estimateurs finaux soient plus précis. Des preuves numériques montrent l'efficacité du réticissement double appliqué aux estimations sur petits domaines de la moyenne si l'on prend pour critère l'EQM.

Une autre contribution importante du présent article est l'obtention d'intervalles de confiance pour les moyennes de petits domaines. La littérature relative à l'estimation sur petits domaines traite avant tout des estimations ponctuelles et de leurs erreurs-types ; pourtant, il est bien connu que la pratique classique consistant à utiliser l'estimation ponctuelle $\pm q \times \text{erreur-type}$, où q est la valeur seuil Z (normale standard) ou t , ne produit pas des probabilités de couverture exactes des intervalles ; voir Hall et Maiti (2006) et Chatterjee, Lahiri et Li (2008) pour plus de précisions. Les travaux antérieurs, qui sont fondés sur la technique du bootstrap, sont d'un usage limité en raison de l'estimation

Estimation des intervalles de confiance des paramètres de petit domaine avec rétrécissement des moyennes et des variances

Sarat C. Dass, Tapabrata Maiti, Hao Ren et Samiran Sinha¹

Résumé

Nous proposons une nouvelle approche d'estimation sur petits domaines fondée sur la modélisation conjointe des moyennes et des variances. Le modèle et la méthodologie que nous proposons améliorent non seulement les estimateurs sur petits domaines, mais donnent aussi des estimateurs « lissés » des vraies variances d'échantillonnage. Le maximum de vraisemblance des paramètres du modèle est estimé au moyen de l'algorithme EM en raison de la forme non classique de la fonction de vraisemblance. Les intervalles de confiance des paramètres de petit domaine sont obtenus en adoptant une approche de la théorie de la décision plus générale que l'approche classique de minimisation de la perte quadratique. Les méthodes concurrentes proposées dans la littérature. Une justification théorique des propriétés effectives des estimateurs et intervalles de confiance résultants est également présentée.

Mots clés : Algorithme EM ; Bayes empirique ; modèles hiérarchiques ; échantillonnage réjectif ; variance d'échantillonnage ; estimation sur petits domaines.

1. Introduction

L'estimation sur petits domaines et les techniques statistiques qui s'y rapportent sont des sujets qui ont fait l'objet d'une attention croissante ces dernières années. De nombreux organismes, tant publics que privés, cherchent à obtenir des estimations sur petits domaines fiables pour prendre des décisions stratégiques utiles. La surveillance de la situation socioéconomique et de l'état de santé de divers groupes définis selon l'âge, le sexe et la race pour lesquels s'observeraient des tendances distinctes sur de petites régions géographiques est un exemple d'application pratique des techniques d'estimation sur petits domaines.

Il est aujourd'hui généralement reconnu que les estimations directes à partir de données d'enquête calculées pour les petits domaines ne sont d'ordinaire pas fiables parce que leurs erreurs-types et coefficients de variation sont très souvent grands. Il devient donc nécessaire d'obtenir de meilleures estimations, d'une plus grande précision. Des approches fondées explicitement ou implicitement sur un modèle sont élaborées pour relier des petits domaines et obtenir une plus grande précision par « emprunt d'information » à des domaines similaires. Cette technique d'estimation est également appelée estimation par rétrécissement, ou estimation à rétrécisseur, puisque les estimations directes sont « rétrécies » afin qu'elles se rapprochent de la moyenne globale. Les estimations directes d'après les données d'enquête et les variances d'échantillon sont les principaux ingrédients qui entrent dans la création des modèles d'estimation sur petits domaines de niveau agrégé. La stratégie de modélisation repose habituellement sur l'hypothèse que les variances d'échantillonnage sont connues, tandis qu'un

modèle de régression linéaire approprié est utilisé pour les moyennes. Pour des renseignements détaillés sur ces développements, le lecteur est invité à consulter Ghosh et Rao (1994), Pfeffermann (2002) et Rao (2003). Les modèles habituels au niveau du domaine suscitent deux critiques importantes. Premièrement, en pratique, les variances d'échantillonnage sont des quantités estimées qui sont donc sujettes à d'importantes erreurs. Il en est ainsi parce qu'elles sont souvent fondées sur des tailles d'échantillon équivalentes à celles qui servent au calcul des estimations directes. Deuxièmement, en raison de l'hypothèse que les variances d'échantillonnage sont connues et fixes formulée dans les modèles d'estimation sur petits domaines classiques, l'incertitude que comporte l'estimation de la variance n'est pas prise en compte dans la stratégie d'inférence globale.

Des tentatives en vue de modéliser uniquement les variances d'échantillonnage ont été faites antérieurement ; voir, par exemple, Maples, Bell et Huang (2009), Gershunskaya et Lahiri (2005), Hufn, Elling et Gershunskaya (2002), Cho, Elling, Gershunskaya et Hufn (2002), Valliant (1987), et Otto et Bell (1995). Dans leurs articles, Wang et Fuller (2003) et Rivest et Vandal (2003) ont étendu l'estimation de l'erreur quadratique moyenne (EQM) asymptotique des estimateurs sur petits domaines au cas où l'on estime les variances d'échantillonnage au lieu de s'appuyer sur l'hypothèse classique que les variances sont connues. En outre, You et Chapman (2006) ont considéré la modélisation des variances d'échantillonnage avec inférence en appliquant des techniques d'estimation entièrement bayésiennes. De nombreux praticiens ont jugé nécessaire de modéliser la variance. Les progrès les plus récents dans ce domaine

1. Sarat C. Dass et Tapabrata Maiti, Department of Statistics & Probability, Michigan State University, Courtiel : maiti@stat.msu.edu ; Hao Ren, CTB/McGraw-Hill, 20 Ryan Ranch Rd, Monterey, CA 93940 ; Samiran Sinha, Department of Statistics, Texas A & M University.

d'exacétude. Pour être efficaces, les interventions pour prévenir la non-réponse dans les études longitudinales doivent être ciblées sur les cas les moins susceptibles de répondre, parce que ceux-ci sont probablement ceux qui diffèrent le plus des répondants et, par conséquent, constituent la source principale de biais. C'est dans cette situation que la méthode des courbes ROC peut être particulièrement utile, parce que, comme le montre Swets, Dawes et Monahan (2000), il est possible de déterminer le seuil optimal pour le score de propension à répondre fondé sur les coûts et les avantages de l'intervention d'après les taux de vrais et de faux positifs qu'implique le seuil. Une évaluation plus détaillée de ces questions dépasse le cadre du présent article, mais comprendrait l'examen d'interventions pour prévenir différents types de non-réponse, et les avantages des réductions éventuelles du biais et de la variabilité découlant d'un échantillon de plus grande taille et dont les caractéristiques sont plus proches de celles de l'échantillon cible.

Remerciements

La présente étude a été financée par l'Economic and Social Research Council du Royaume-Uni dans le cadre de la Survey Design and Measurement Initiative (réf. RES-175-25-0010).

Bibliographie

Copas, J. (1999). The effectiveness of risk scores: The logit rank plot. *Applied Statistics*, 48, 165-183.

Groves, R.M. (2006). Nonresponse rates and non-response bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Swets, J.A., Dawes, R.M. et Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Sciences in the Public Interest*, 1, 1-26.

Plewits, I., Ketende, S.C., Joshi, H. et Hughes, G. (2008). The contribution of residential mobility to sample loss in a birth cohort study: Evidence from the first two waves of the Millennium Cohort Study. *Journal of Official Statistics*, 24, 365-385.

Plewits, I. (Ed.) (2007b). *The Millennium Cohort Study: Technical Report on Sampling* (4^e Ed.). Londres : Institute of Education, University of London.

Plewits, I. (2007a). Non-response in a birth cohort study: The case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, 10, 325-334.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford : OUP.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2^e Ed.). New York : John Wiley & Sons, Inc.

Lepkowski, J.M., et Couper, M.P. (2002). Nonresponse in the second wave of longitudinal household surveys. Dans *survey Nonresponse*, (Eds., R.M. Groves et coll.). New York : John Wiley & Sons, Inc.

Krizanowski, W.J., et Hand, D.J. (2009). *ROC Curves for Continuous Data*. Boca Raton, FL : Chapman and Hall/CRC.

Hawkes, D., et Plewits, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society A*, 169, 479-491.

Harrell, F.E. Jr., Lee, K.L. et Mark, D.B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361-387.

Tableau 2
Estimation de l'exactitude pour les modèles de propension à répondre améliorés, deuxième vague de la MCS

Mesure de l'exactitude	Non-réponse globale ⁽¹⁾	Type de non-réponse	Raison de la non-réponse
ASC	0,70	Non-réponse à une vague ⁽²⁾	0,71
Gini	0,41		0,41
Courbe logit sur rangs : pente	0,47		0,46
Taille de l'échantillon	18 148		16 745
(1) Inclut le consentement (Rapport de cotes (RC) = 2,1, e.-l. = 0,20) et le vote (RC = 1,4, e.-l. = 0,08).			
(2) Inclut le vote seulement (RC = 1,4, e.-l. = 0,11), consentement pas important ($t = 1,33$; $p > 0,18$).			
(3) Inclut le consentement (RC = 2,7, e.-l. = 0,26) et le vote (RC = 1,4, e.-l. = 0,09).			
(4) Inclut le consentement (RC = 2,6, e.-l. = 0,32), le vote (RC = 1,3, e.-l. = 0,10) et le score du quartier (RC = 1,02, e.-l. = 0,014).			
(5) Inclut le consentement (RC = 1,6, e.-l. = 0,20) et le vote (RC = 1,5, e.-l. = 0,11).			

Tableau 3
Transitions d'emploi pondérées (erreurs-types), deuxième vague de la MCS

Variable	Poids de sondage seulement	Poids global ⁽¹⁾	Poids global ⁽²⁾
Pas de changement	0,30 (0,0053)	0,30 (0,0056)	0,31 (0,0056)
Emploi → pas d'emploi	0,34 (0,0059)	0,35 (0,0059)	0,35 (0,0060)
Pas d'emploi → emploi	0,37 (0,0073)	0,35 (0,0073)	0,35 (0,0073)
Étendue du poids ⁽³⁾	0,23 – 2,0	0,19 – 4,1	0,19 – 6,3
Taille de l'échantillon	14 891	14 796	14 733

(1) Fondé sur le produit des poids de sondage et des poids de non-réponse en utilisant le modèle qui sous-tend le tableau 1.
(2) Poids de non-réponse fondé sur un modèle qui inclut les variables de consentement et de vote.
(3) Tous les poids sont normalisés de manière que leur moyenne soit égale à un.

5. Discussion

Les méthodologistes d'enquête qui travaillent avec des données longitudinales sont confrontés depuis longtemps au problème de non-réponse. Presque toutes les études longitudinales souffrent d'une accumulation des cas de non-réponse au fil du temps. Il est fréquent, même pour des études bien réalisées et bien établies, d'obtenir des données pour moins de la moitié de l'échantillon cible. Par ailleurs, il est possible d'en savoir beaucoup sur les corrélats de différents types de non-réponse en s'appuyant sur les variables auxiliaires provenant de vagues antérieures. L'objectif principal du présent article était de présenter un moyen différent de réfléchir à l'utilité des approches qui s'appuient sur des modèles linéaires généraux à la fois pour construire des pondérations par l'inverse des probabilités et pour faciliter les imputations. Traiter les prédicteurs linéaires issus des modèles de régression comme des scores de propension à répondre, puis créer de courtes ROC offre des méthodes pour résumer l'information contenue dans ces scores afin de l'utiliser pour évaluer l'exactitude de la discrimination et de la prédiction pour différents types de non-réponse.

Une des implications de cette constatation est que certains cas de non-réponse découlent de facteurs circonstanciels, sans importance pris individuellement, qui peuvent raisonnablement être considérés comme le hasard. Notre étude étaye dans une certaine mesure cette hypothèse en ce sens que l'exactitude des modèles pour la non-réponse globale, la non-réponse à une vague et les autres cas non produits (les deux dernières catégories étant reliées) a été peu modifiée par l'introduction des variables de vote et de consentement. Par ailleurs, ces variables (et le score de conditions du quartier) ont amélioré le pouvoir de discrimination entre les cas produits, d'une part, et les cas d'attrition et de refus (qui sont aussi reliés), d'autre part. Néanmoins, le pouvoir de discrimination pour ces deux catégories demeurerait plus faible que pour les autres types de non-réponse. Une deuxième implication éventuelle est que les modèles ne réalisent pas bien la discrimination parce que l'on a affaire à une répartition des données manquantes qui n'est pas due au hasard (NMAR pour *not missing at random*) au sens de Little et Rubin (2002). Autrement dit, il pourrait se produire après la vague précédente des changements de circonstances qui influencent la non-réponse durant la vague courante. Les implications de nos constatations en ce qui concerne la prédiction sont qu'il pourrait être difficile de prédire quels cas deviendront des non-répondants avec un haut degré

Tableau 1
Estimations de l'exactitude d'après les modèles de propension à répondre, deuxième vague de la MCS

Mesure de l'exactitude	Non-réponse globale ⁽²⁾	Type de non-réponse ⁽²⁾	Raison de la non-réponse ⁽²⁾		
	Non-réponse à une vague	Attrition	Refus	Autres cas non productifs	
ASC ⁽¹⁾	0,69	0,71	0,69	0,68	0,77
Gini ⁽¹⁾	0,38	0,42	0,39	0,40	0,53
Courbe logit sur rangs : pente ⁽¹⁾	0,45	0,51	0,44	0,40	0,63
Taille de l'échantillon	18 230	16 210	16 821	16 543	16 513

(1) ASC estimée sous l'hypothèse binomiale (Krzyszowski et Hand 2009) : limites de confiance à 95 % pour a) ASC n'excédant pas 0,015, b) coefficient de Gini et pente de la courbe logit sur rangs n'excédant pas 0,03.
(2) Fondé sur une régression logistique, avec prise en compte du plan de sondage en utilisant les commandes svy de STATA avec la taille d'échantillon correspondant à la somme des cas productifs et des cas de non-réponse selon la catégorie.

La spécification correcte des modèles destinés à expliquer la non-réponse peut être difficile. De nouveaux candidats susceptibles d'être inclus dans un modèle peuvent apparaître après que l'on ait estimé ce dernier ainsi que les pondérations correspondantes par l'inverse de la probabilité, tandis que d'autres demeurent inconnus. Quelle est l'importance de l'effet que pourrait avoir sur les mesures de l'exactitude l'inclusion de nouvelles variables ? Ici, nous examinons les effets de l'ajout de trois nouvelles variables aux modèles de la MCS : i) le fait que les répondants donnent ou non leur consentement pour que leurs réponses à l'enquête soient apparées aux dossiers de santé à la première vague ; ii) un score de conditions dans le quartier calculé d'après les observations faites par l'intervieweur à la deuxième vague, et iii) le fait que, à la première vague, le répondant principal a ou non déclaré avoir voté aux dernières élections générales au Royaume-Uni. Les deux premières de ces variables n'étaient pas disponibles pour les analyses résumées au tableau 1 : le refus du consentement à la vague t pourrait être suivi par un refus global à la vague $t + 1$, et la non-réponse pourrait être plus fréquente dans les quartiers pauvres. La variable de vote est un indicateur de l'engagement social qui pourrait être associé à la probabilité de répondre. Comme le score des conditions dans le quartier n'a pas pu être obtenu pour les cas qui n'ont pas été localisés, nous utilisons cette variable uniquement dans le modèle où sont comparés les cas de refus aux cas productifs.

4.2 Utilisation des pondérations pour corriger la non-réponse

Bien que la non-réponse à la deuxième vague de la MCS soit systématiquement reliée à un certain nombre de variables mesurées durant la première vague ou après, nous avons constaté que la capacité du modèle à faire la distinction entre les catégories de non-réponse et à prédire ces dernières n'est pas très grande. Nous allons maintenant considérer l'effet que les pondérations produites à partir des modèles de propension à répondre ont sur une estimation longitudinale d'intérêt. Nous nous concentrons sur les tan-sitions entre l'absence d'emploi et la possession d'un em-ploi entre les deux vagues. Comme le soutient Groves (2006), la clé en vue de résoudre le problème de biais causé par les données manquantes consiste à trouver des variables qui prédisent de ces variables prédisant les données man-lesquelles sont également reliées à la variable d'intérêt. Nous constatons que toutes les variables qui prédisent la non-réponse globale sont également associées au fait que la réponse globale vague, conditionnellement au fait qu'elle travail-le ou ne travaille pas à la deuxième vague, conditionnellement au fait qu'elle travail-le ou ne travaille pas à la première vague, de sorte que nous devrions nous attendre à ce que l'application des poids de non-réponse ré-duise le biais. Les résultats, présentés au tableau 3, montrent que, comparativement à l'utilisation des poids de sondage, l'ajout des poids de non-réponse fondés sur le modèle qui sous-tend le tableau 1 produit de petites corrections des probabilités de transition estimées. Par contre, les variables de consentement et de vote n'ont aucun effet supplémentaire, ce qui est en harmonie avec l'augmentation marginale de l'exactitude indiquée au tableau 2.

7 ans. À la deuxième vague, 19 % de l'échantillon cible, qui n'inclut pas les enfants décédés et émigrés, ont été non productifs. Les cas non productifs étaient répartis de manière égale entre les cas de non-réponse à la vague et les cas d'attrition, et entre les cas de refus et les autres cas non productifs (pas localisés, pas contactés, etc.).

4. Analyse de la non-réponse

4.1 Exactitude de la discrimination et de la prédiction

Plewis (2007a) et Plewis et coll. (2008) montrent que les variables mesurées durant la première vague de la MCS qui sont associées à l'attrition durant la deuxième vague ne sont pas nécessairement associées à la non-réponse à une vague à ce moment-là (et inversement). Il en est de même des corrélats du refus et des autres cas non productifs. Le tableau 1 donne les estimations de l'exactitude d'après les modèles de propension à répondre. L'estimation du coefficient de Gini pour la non-réponse globale (0,38) est relativement faible : elle correspond à une ASC de 0,69 qui est la probabilité d'attribuer correctement (en se fondant sur leurs probabilités prédites) une paire de cas (un répondant et un non-répondant), ce qui indique que la discrimination entre les non-répondants et les répondants d'après les scores de propension à répondre n'est pas particulièrement bonne. Elle est légèrement meilleure pour les non-répondants à une vague que pour les cas d'attrition, et est nettement meilleure pour les autres cas non productifs que pour les refus. Ces estimations ont été obtenues au moyen de comparaisons par paires de chaque catégorie de non-réponse avec le fait d'être un répondant. Un tableau comparable se dégage lorsque l'on examine les pentes des courbes logit sur rangs, quoique celles-ci fassent ressortir plus clairement les différences de prédictivité pour les différents types de non-réponse et pour les raisons de la non-réponse.

3. La Millennium Cohort Study

La mesure du pouvoir prédictif du score de propension. Selon Copas, la pente est plus sensible aux changements de spécification du modèle de propension à répondre et à ceux de la prévalence du résultat que ne l'est le coefficient de Gini. Une bonne estimation de la pente peut être obtenue en calculant les quantiles de la variable sur les axes des y et des x , puis en ajustant un simple modèle de régression. La mesure dans laquelle les scores de propension à répondre permettent de distinguer les répondants des non-répondants est un indicateur de l'efficacité de tout ajustement statistique pour tenir compte des données manquantes. Un manque de pouvoir de discrimination donne à penser que des prédictifs importants manquent dans le score de propension à répondre ou qu'une part importante du processus qui dicte l'existence des données manquantes est essentiellement aléatoire. La mesure dans laquelle les scores de propension à répondre prédisent si un cas sera un non-répondant aux vagues subséquentes – et de quel type de non-répondant il s'agira – est un indice du succès qu'aura toute intervention destinée à réduire la non-réponse.

L'échantillon de la première vague de la Millennium Cohort Study (MCS) réalisée au Royaume-Uni comprend 18 552 familles dans lesquelles est né un enfant au cours d'une période de 12 mois durant les années 2000 et 2001 et qui vivaient dans des circonscriptions électorales choisies du Royaume-Uni au moment où l'enfant était âgé de 9 mois. Le taux initial de réponse était de 72 %. Les secteurs où les proportions de familles noires et asiatiques sont élevées, les secteurs défavorisés et les trois plus petits pays du Royaume-Uni sont tous surreprésentés dans l'échantillon qui est structuré en grappes et stratifié de façon disproportionnée, comme l'a décrit Plewis (2007b). Les quatre premières vagues ont eu lieu lorsque les enfants membres de la cohorte étaient âgés (environ) de 9 mois, 3 ans, 5 ans et

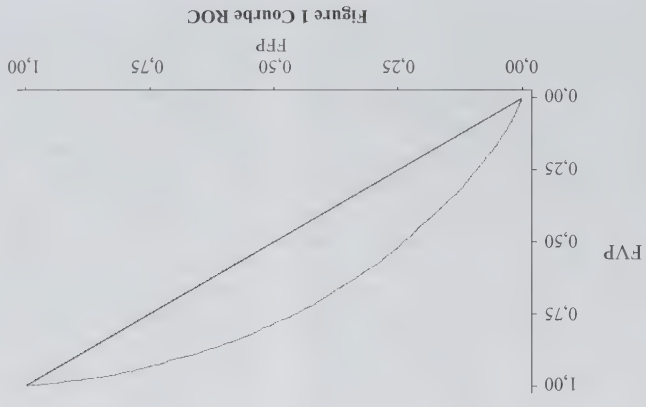


Figure 1 Courbe ROC

non-répondant étant donné un score de propension supérieur ou inférieur au seuil.

De manière plus formelle, soit D et \underline{D} la présence et l'absence du mauvais résultat (c'est-à-dire la non-réponse) et définissons $+$ ($s > c$) et $-$ ($s \leq c$) comme étant les tests positif et négatif dérivés du score de propension à répondre et de son seuil. Alors, pour la discrimination, nous nous intéressons à $P(+|D)$, la fraction de vrais positifs (FVP) ou sensibilité du test, et à $P(-|\underline{D})$, sa spécificité, qui est égale à un moins la fraction de faux positifs (1 - FFP). Pour la prédiction, par contre, nous nous intéressons à $P(D|+)$, la valeur prédictive positive (VPP) et à $P(\underline{D}|-)$, la valeur prédictive négative (VPN). Si la probabilité d'un test positif ($P(+)=\tau$) est la même que la prévalence du mauvais résultat ($P(D)=p$), les inférences au sujet de la discrimination et de la prédiction sont essentiellement les mêmes : la sensibilité est égale à la VPP et la spécificité est égale à la VPN. Cependant, généralement, {FVP, FFP, p } and {VPP, VPN, τ } communiquent des éléments d'information différents. La FVP peut être représentée graphiquement en fonction de la FFP pour tout seuil de score de risque c . On obtient ainsi la courbe de la fonction d'efficacité du receveur, ou courbe ROC (figure 1). Kizanowski et Hand (2009) discutent en détail de la façon d'estimer les courbes ROC. L'aire sous la courbe (ASC) – l'aire délimitée par la courbe ROC et par l'axe des x dans la figure 1 – est particulièrement intéressante et sa valeur peut varier de 1 (discrimination parfaite) à 0,5, c'est-à-dire l'aire sous la diagonale, ce qui implique l'absence de discrimination). L'ASC peut être interprétée comme la probabilité d'attribuer une paire de cas, un répondant et un non-répondant, aux catégories correctes, en se souvenant qu'une réponse devinée correspondrait à une probabilité de 0,5. Une transformation linéaire de l'ASC ($=2*ASC-1$) – parfois appelée coefficient de Gini et équivalente à l'indice de corrélation de rang D de Somer (Harrell, Lee et Mark 1996) – est fréquemment utilisée comme mesure plus naturelle que l'ASC, parce que sa valeur varie de 0 à 1.

Copas (1999) propose la courbe de type logit sur rangs comme alternative à la courbe ROC pour évaluer le pouvoir prédictif d'un score de propension. Si le score de propension est calculé d'après une régression logistique, la courbe logit sur rangs est simplement la représentation graphique du prédicteur linéaire issu du modèle en fonction de la transformation logarithmique du rang proportionnel du score de propension. Plus généralement, il s'agit d'une représentation graphique de $\logit(p'_i)$, où p'_i est la probabilité estimée d'après toute forme de (1), c'est-à-dire $p(D|x, x^*, z)$, en fonction des logits des rangs proportionnels (r/n) où r est le rang du cas i ($i=1, \dots, m$) sur le score linéaire et sa pente – qui peut varier de zéro à un – est une

x^{pi} représente les caractéristiques fixes du sujet i mesurées à la vague un, $p=0, \dots, P; x_0=1$ pour tout i ;

x^{q-i-k} représente les caractéristiques variant avec le temps du sujet i , mesurées aux vagues $t-k, q=1, 2, \dots, \text{soit } k \text{ est égal à } 1$;

z^{r-i-k} représente les caractéristiques variant avec le temps du processus de collecte des données, mesurées pour le sujet i aux vagues $t-k, r=1, \dots, R, k=0, 1, \dots, \text{soit } k \text{ est égal à } 1$, mais peut être égal à 0 pour des variables telles que le nombre de contacts avant d'obtenir une réponse.

Le modèle (1) peut être étendu facilement à plus de deux catégories de réponse, comme {réponse, non-réponse à une vague, attention}. D'autres approches sont également possibles. Par exemple, il est souvent plus commode de modéliser la probabilité de ne pas répondre seulement à la vague $t=t^*$ en ce qui a trait aux variables mesurées durant les vagues antérieures $t^*-k, k \geq 1$ ou, en l'absence de non-réponse à une vague de sorte que la courbe de non-réponse est monotone plutôt qu'arbitraire, de modéliser le temps écoulé jusqu'à l'attrition comme un processus de survie.

Les probabilités de réponses estimées p_i pour $t=t^*$, sont dérivées des probabilités de non-réponse estimées en (1) et peuvent servir à produire des pondérations égales à l'inverse de la probabilité $g_i(=1/p_i)$. Ces pondérations sont très souvent utilisées (voir la section 4.2 pour un exemple) pour corriger le biais dû à la non-réponse sous l'hypothèse que les données manquantes sont dues au hasard (MAR pour *missing at random*), comme l'ont défini Little et Rubin (2002).

2.1 Évaluation de l'exacitude des prédictions

Une méthode très répandue pour évaluer l'exacitude des modèles tels que (1) consiste à estimer leur adéquation en utilisant une ou plusieurs statistiques pseudo- R^2 possibles. Les estimations du pseudo- R^2 ne sont pas particulièrement utiles dans le présent contexte, en partie parce qu'elles sont difficiles à comparer entre jeux de données, mais aussi parce qu'elles évaluent l'adéquation globale du modèle et ne font donc pas la distinction entre l'exacitude du modèle pour les répondants et pour les non-répondants pris séparément. Comme le souligne Pepe (2003), l'exacitude possède deux composantes apparentes : la discrimination (ou classification) et la prédiction. Par discrimination, on entend les probabilités conditionnelles d'avoir un score de propension à répondre (s) : le prédicteur linéaire provenant de (1) à répondre à un seul choix (c) sachant qu'une personne est un non-répondant. Par ailleurs, par prédiction, on entend les probabilités conditionnelles de devenir un

Evaluation de l'exactitude des modèles de propension à répondre dans les études longitudinales

Ian Plewis, Sosthenes Ketende et Lisa Calderwood

Résumé

La question de la non-réponse dans les études longitudinales est abordée en évaluant l'exactitude des modèles de propension à répondre construits pour distinguer et prédire les divers types de non-réponses. Une attention particulière est accordée aux mesures sommaires dérivées des courbes de la fonction d'efficacité du receveur, ou courbes ROC (de l'anglais *receiver operating characteristics*), ainsi que des courbes de type logit sur rangs. Les concepts sont appliqués à des données provenant de la Millennium Cohort Study du Royaume-Uni. Selon les résultats, la capacité de faire la distinction entre les divers types de non-répondants et de les prévoir n'est pas grande. Les poids produits au moyen des modèles de propension à répondre ne donnent lieu qu'à de faibles corrections des transitions entre situations d'emploi. Des conclusions sont tirées quant aux possibilités d'intervention en vue de prévenir la non-réponse.

Mots clés : Études longitudinales ; données manquantes ; pondération ; scores de propension ; courbes ROC ; Millennium Cohort Study.

1. Introduction

Les exemples d'études ayant adopté les prédicteurs des différents types de non-réponse et des raisons de la non-réponse dans les études longitudinales sont nombreux. De telles modélisations ont été rendues possibles grâce à la capacité de se servir de variables auxiliaires pour lesquelles des données ont été obtenues auprès des membres de l'échantillon avant (et après) les cycles auxquels ces membres non pas répondu. Notamment, Lepkowski et Couper (2002) proposent une analyse séparant les cas de refus des cas dont les répondants n'ont pu être localisés ou contacts : Hawkes et Plewis (2006) font la distinction entre les non-répondants à une vague et les cas d'attrition dans la National Child Development Study menée au Royaume-Uni ; et Plewis (2007a) et Plewis, Ketende, Joshi et Hughes (2008) examinent la non-réponse aux deux premières vagues de la Millennium Cohort Study du Royaume-Uni. Le présent article porte sur la façon dont nous pouvons évaluer l'exactitude de ces modèles de propension à répondre (Little et Rubin 2002). Il s'appuie sur un cadre fréquemment utilisé en épidémiologie (Pepé 2003) et en criminologie (Copas 1999) pour évaluer les scores de risque, mais qui, autant que nous sachions, n'a pas été utilisé auparavant dans le domaine de la recherche des enquêtes. Les modèles de propension à répondre peuvent être utilisés pour construire des poids destinés à éliminer les biais des estimations, à faciliter les imputations, et à prédire les non-répondants aux vagues futures afin d'orienter les ressources pour le travail sur le terrain vers ces répondants qui, autrement, pourraient être perdus. Toutefois, l'exactitude des modèles de propension à répondre n'a pas reçu l'attention

qu'elle aurait dû en ce qui concerne la capacité à faire la distinction entre les répondants et les non-répondants et à prédire la non-réponse. De bonnes estimations de l'exactitude peuvent être utilisées pour comparer l'efficacité de différents modèles de pondération et pour faciliter la répartition des ressources limitées réservées au travail sur le terrain afin de réduire la non-réponse. La présentation de l'article est la suivante. Le cadre d'évaluation de l'exactitude est exposé à la section suivante. À la section 3, nous présentons la Millennium Cohort Study du Royaume-Uni et à la section 4, nous illustrons les méthodes au moyen de données provenant de cette étude. Enfin, à la section 5, nous présentons nos conclusions.

2. Modèles pour la prédiction de la non-réponse

Un modèle type de propension à répondre pour un résultat binaire (par exemple Hawkes et Plewis 2006) est donné par :

$$f(\pi_n) = \sum^p \beta^p x^{pi} + \sum^q \gamma^q x^{qi,1-k} + \sum^k \delta^k x^{ki,1-k} \quad (1)$$

où

- $\pi_n = E(r_n)$ est la probabilité que le sujet i ne réponde pas à la vague t ; $r_n = 0$ pour une réponse et 1 pour une non-réponse ; f est une fonction appropriée, telle que la fonction logit ou probit ;
- $i = 1, \dots, n$ où n est la taille de l'échantillon observé à la première vague ;
- $t = 1, \dots, T$ où T est le nombre de vagues pour lesquelles r_n est enregistré pour le sujet i ;

Remerciements

Les travaux de recherche ont été financés en partie par une entente de coopération entre le Natural Resources Conservation Service du US Department of Agriculture et la Iowa State University. Les auteurs remercient F. Jay Breidt, trois examinateurs anonymes et le rédacteur associé de leurs commentaires constructifs.

Bibliographie

Cao, W., Tsaias, A.A. et Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96, 723-734.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34, 305-334.

Da Silva, D.N., et Opsomer, J.D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *Canadian Journal of Statistics*, 34, 563-579.

Da Silva, D.N., et Opsomer, J.D. (2009). Pondération par la propension à répondre non paramétrique fondée sur la régression par polyômes locaux pour corriger la non-réponse aux enquêtes. *Techniques d'enquête*, 35, 2, 179-192.

Duncan, K.B., et Stasny, E.A. (2001). Utilisation de scores de propension pour contrôler le biais de couverture dans les enquêtes téléphoniques. *Techniques d'enquête*, 27, 2, 131-141.

Durand, G.B., et Skinner, C. (2006). Utilisation de méthodes de traitement des données manquantes pour corriger l'erreur de mesure dans une fonction de distribution. *Techniques d'enquête*, 32, 1, 27-39.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section, American Statistical Association*, 227-232.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistics Section, American Statistical Association*, 197-202.

Fuller, W.A., Loughin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la *d'enquête*, 20, 1, 79-89.

Iannacchione, V.G., Milne, J.G. et Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 637-642.

Isak, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Kim, J.K. (2004). Finite sample properties of multiple imputation estimators. *The Annals of Statistics*, 32, 766-783.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Kim, J.K., et Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35, 501-514.

Kim, J.K., et Rao, J.N.K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22, 329-349.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-296.

Neyo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business and Economic Statistics*, 21, 43-52.

Pfeffermann, D., Krieger, A.M. et Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Randles, R.H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, 10, 462-474.

Rizzo, L., Kailton, G. et Brck, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 1, 43-53.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387-394.

Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Singh, A.C., et Folsom, R.E. (2000). Bias corrected estimating function approach for variance estimation adjusted for poststratification. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 610-615.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Pour calculer le score de propension, nous avons postulé un modèle de réponse de la forme

$$p(x; \phi) = \frac{\exp(\phi_0 + \phi_1 x)}{1 + \exp(\phi_0 + \phi_1 x)}$$

pour estimer les paramètres. Pour obtenir l'estimateur ajusté sur le score de propension augmenté, nous avons postulé pour la variable d'intérêt un modèle de la forme

$$m(x; \beta) = \frac{1 + \exp(\beta_0 + \beta_1 x)}{\exp(\beta_0 + \beta_1 x)} \tag{32}$$

Donc, le modèle (32) est un modèle vrai sous (Pop1), mais non sous (Pop2).

Nous avons calculé quatre estimateurs :

- 1. (ASP-EMV) : estimateur ajusté sur le score de propension (3) en utilisant l'estimateur du maximum de vraisemblance de ϕ .

- 2. (ASP-CAL) : estimateur ajusté sur le score de propension (3) avec \hat{p}_i satisfaisant la contrainte de calage (15) sur $(1, x)$.
- 3. (AUG-1) : estimateur ajusté sur le score de propension augmenté $\hat{\theta}_{ASP}^{(19)}$ avec $\hat{\beta}$ calculé par la méthode du maximum de vraisemblance.

- 4. (AUG-2) : estimateur ajusté sur le score de propension augmenté $\hat{\theta}_{ASP}^{(19)}$ avec $\hat{\beta}$ calculé par la méthode de

Cao et coll. (2009) discutée à la remarque 1.

Nous avons considéré l'estimateur ajusté sur le score de propension augmenté (19) avec $\hat{p}_i = p_i(\phi)$, où ϕ est l'estimateur du maximum de vraisemblance de ϕ . Le premier estimateur ajusté sur le score de propension augmenté (AUG-1) utilisait $\hat{m}_i = m(x_i; \beta)$ avec $\hat{\beta}$ obtenu par résolution de $\sum_{h=1}^4 \sum_{i \in A_h} w_{hi} r_{hi} \{V_{hi} - m(x_{hi}; \beta)\} (1, x_{hi}) = 0$, où A_h est l'ensemble d'indices figurant dans l'échantillon de la strate h et w_{hi} est le poids d'échantillonage de l'unité i dans la strate h .

Le tableau 2 donne les résultats de simulation pour chaque méthode. Pour chaque population, l'estimateur ajusté sur le score de propension augmenté montre une certaine amélioration de la variance comparativement à

6. Conclusion

Nous avons considéré le problème de l'estimation de la moyenne de y en population finie en présence de non-

réponse en utilisant la méthode du score de propension. Nous avons calculé le score de propension à l'aide d'un modèle paramétrique de la probabilité de réponse et discuté de certaines propriétés asymptotiques des estimateurs ajustés sur le score de propension. En particulier, l'estimateur ajusté sur le score de propension optimal est établi en émettant une hypothèse supplémentaire au sujet de la distribution de y . Le score de propension pour l'estimateur ajusté sur le score de propension optimal peut-être obtenu à l'aide du modèle de score de propension augmenté présenté à la section 3. L'estimateur résultant reste convergent, même si le modèle de régression du résultat supposé ne tient pas.

Nous avons limité notre étude au mécanisme de données manquant au hasard dans lequel la probabilité de réponse ne dépend que de x qui est toujours observé. Si le mécanisme de réponse dépend également de y , l'estimation ajustée sur le score de propension devient plus difficile. L'estimation ajustée sur le score de propension quand les données ne manquent pas au hasard dépasse le cadre du présent article et sera le sujet d'une future étude.

Tableau 2
Biais, variance et erreur quadratique moyenne Monte Carlo des quatre estimateurs ponctuels et biais relatifs (BR) en pourcentage et statistique t des estimateurs de variance, fondés sur 5 000 échantillons Monte Carlo

Population	Méthode	$\hat{\theta}_{ASP}$	Variance	EQM	BR (%)	$V(\hat{\theta}_{ASP})$
Pop1	(ASP-EMV)	0,00	0,000750	0,000762	-1,13	-0,57
	(ASP-CAL)	0,00	0,000762	0,000769	-1,45	-0,72
	(AUG-1)	0,00	0,000745	0,000757	-1,73	-0,86
	(AUG-2)	0,00	0,000745	0,000757	-1,83	-0,91
	(ASP-EMV)	0,00	0,000824	0,000826	0,29	0,14
	(ASP-CAL)	0,00	0,000829	0,000835	-0,94	-0,46
Pop2	(ASP-EMV)	0,00	0,000820	0,000823	-0,71	-0,35
	(ASP-CAL)	0,00	0,000822	0,000823	-0,71	-0,35
	(AUG-1)	0,00	0,000820	0,000823	-0,71	-0,35
	(AUG-2)	0,00	0,000820	0,000823	-0,71	-0,35
	(ASP-EMV)	0,00	0,000820	0,000823	-0,71	-0,35
	(ASP-CAL)	0,00	0,000820	0,000823	-0,71	-0,35

3. Les estimateurs de variance sont tous approximativement sans biais. Les estimateurs de variance des estimateurs ajustés sur le score de propension présentent un biais modeste quand la taille de l'échantillon est faible ($n = 100$).

5.2 Deuxième étude

Dans la deuxième étude par simulation, nous avons poursuivi l'étude des estimateurs ajustés sur le score de propension avec un modèle de régression du résultat non linéaire sous un plan de sondage avec probabilités inégales. Nous avons créé deux populations finies stratifiées de (x, y) comprenant quatre strates ($h = 1, 2, 3, 4$), où les x_{hi} étaient des variables indépendantes tirées selon une loi normale $N(1, 1)$ et les y_{hi} étaient des variables dichotomiques prenant la valeur 1 ou 0 tirées selon une loi de Bernoulli de paramètre p_{1yhi} ou p_{2yhi} . Des probabilités différentes ont été utilisées pour ces deux populations, respectivement :

1. Population 1 (Pop1) :

$$p_{1yhi} = 1 / \{1 + \exp(0,5 - 2x)\}.$$

2. Population 2 (Pop2) :

$$p_{2yhi} = 1 / \{1 + \exp\{0,25(x - 1,5)^2 - 1,5\}\}.$$

En plus de x_{hi} et y_{hi} , les variables indicatrices de réponse r_{hi} ont été tirées selon une loi de Bernoulli de paramètre $P_{hi} = 1 / \{1 + \exp(-1,5 + 0,7x_{hi})\}$. Les tailles des quatre strates étaient $N_1 = 1\,000$, $N_2 = 2\,000$, $N_3 = 3\,000$ et $N_4 = 4\,000$, respectivement. Dans chacune des deux populations finies, nous avons procédé au tirage d'un échantillon stratifié de taille $n = 400$ indépendamment sans remise, où un échantillon aléatoire simple de taille $n_h = 100$ a été tiré de chaque strate. Nous avons utilisé $B = 5\,000$ échantillons Monte Carlo dans cette simulation. Le taux de réponse était de 67 % environ.

Tableau 1

Biais, variance et erreur quadratique moyenne (EQM) Monte Carlo des quatre estimateurs ponctuels et biais relatif (BR) en pourcentage et statistique t (stat. t) des estimateurs de variance fondés sur 5 000 échantillons Monte Carlo

n	Méthode	Biais	Variance	EQM	BR(%)	V(%)	Stat. t
100	(ASP-EMV)	-0,01	0,0315	0,0317	-2,34	-1,12	-1,12
	(ASP-CAL)	-0,01	0,0308	0,0309	-3,56	-1,70	-1,70
	(AUG)	0,00	0,0252	0,0252	-0,61	-0,30	-0,30
	(OPT)	0,00	0,0252	0,0252	-0,21	-0,10	-0,10
400	(ASP-EMV)	-0,01	0,00737	0,00746	0,35	0,17	0,17
	(ASP-CAL)	-0,01	0,00724	0,00728	0,29	0,14	0,14
	(AUG)	0,00	0,00612	0,00612	0,07	0,03	0,03
	(OPT)	0,00	0,00612	0,00612	-0,14	-0,07	-0,07

$$V_i = N^{-2} \sum_{j \in A} \sum_{k \in A} \Omega_{jk} \eta_j \eta_k \quad (27)$$

où $\eta_i = \eta_i(\phi, \beta)$ est défini dans (24) avec η_h^* remplacé par η_i ($\sum_{i \in A} w_i \eta_i \mathbf{z}_i(\phi) p_i = \sum_{i \in A} w_i \eta_i^* \mathbf{z}_i(\phi) p_i$), et $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i; \phi, \beta)$. Pour montrer que V_i converge également vers V_i dans (26), il suffit de montrer que $V\{\eta_{\text{HT}} | \mathcal{F}_N, \mathcal{R}_N\} | \mathcal{F}_N = o(1)$, ce qui découle de (25) et de l'existence du moment d'ordre quatre. Voir Kim, Navarro et Fuller (2006). Dans (26), le deuxième terme V_z est

$$V\{E(\eta_{\text{HT}} | \mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\} = V\left(N^{-1} \sum_{i=1}^N \eta_i | \mathcal{F}_N\right) = \frac{1}{N} \sum_{i=1}^N \frac{p_i}{1-p_i} (Y_i - p_i \mathbf{h}_{i*}^* \eta_h^*)^2, \\ \text{ou } \mathbf{h}_i^* = \mathbf{h}(\mathbf{x}_i; \phi_0, \beta^*). \text{ Un estimateur convergent de } V_z \text{ peut s'obtenir sous la forme}$$

$$V_z = \frac{1}{N} \sum_{i=1}^N w_i V_i = \frac{1}{N} \sum_{i=1}^N \frac{p_i^2}{1-p_i} (Y_i - p_i \mathbf{h}_i^* \eta_h^*)^2, \quad (28) \\ \text{où } \eta_h^* \text{ est défini d'après (27). Donc,}$$

$$V(\theta^{\text{asp}}) = V^1 + V_z^2 \quad (29)$$

est un estimateur convergent de la variance de l'estimateur ajusté sur le score de propension défini dans (3) avec $\hat{p}_i = p_i(\phi)$ satisfaisant (6), où V^1 est donné par (27) et V_z^2 par (28).

Notons que le premier terme de la variance totale est $V^1 = O_p(n^{-1})$, mais que le deuxième terme est $V_z^2 = O_p(N^{-1})$. Donc, quand la fraction d'échantillonnage nN^{-1} est négligeable, c'est-à-dire $nN^{-1} = o(1)$, le deuxième terme V_z^2 peut être ignoré et V^1 est un estimateur convergent de la variance totale. Sinon, il faut tenir compte du deuxième terme V_z^2 pour pouvoir construire un estimateur de variance convergent comme dans (29).

Remarque 2 L'estimation de la variance de l'estimateur ajusté sur le score de propension optimal sous le modèle de score de propension augmenté (18) avec $(\hat{\phi}, \hat{\lambda})$ satisfaisant (20) peut être dérivé de (29) en utilisant $\eta_i = b_0 + b_1 m_i + \eta_{h2}^* p_i \mathbf{h}_i + r_1 p_i^2 (\mathbf{Y}_i - b_0 - b_1 m_i - \eta_{h2}^* p_i \mathbf{h}_i)$ où (b_0, b_1) et η_{h2}^* sont définis dans (21) et (22), respectivement.

5. Études par simulation

5.1 Première étude

Deux études par simulation ont été exécutées pour première simulation, nous avons créé une population finie

de taille $N = 10\,000$ à partir de la loi normale multivariée suivante :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim N \left(\begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right).$$

La variable d'intérêt y a été construite sous la forme $y = 1 + x_1 + e$. Nous avons également généré des variables indicatrices de réponse r_i indépendamment à partir d'une loi de Bernoulli de paramètre

$$p_i = \frac{\exp(2 + x_{z1})}{\exp(2 + x_{z1}) + 1}.$$

Partant de la population finie, nous avons utilisé l'échantillonnage aléatoire simple pour sélectionner deux échantillons de taille $n = 100$ et $n = 400$, respectivement. Nous avons utilisé $B = 5\,000$ échantillons Monte Carlo dans la simulation. Le taux de réponse moyen était de 69,6 % environ. Pour calculer le score de propension, nous avons postulé un modèle de réponse de la forme

$$p(\mathbf{x}; \phi) = \frac{\exp(\phi_0 + \phi_1 x_2)}{1 + \exp(\phi_0 + \phi_1 x_2)} \quad (30)$$

et un modèle de régression du résultat de la forme

$$m(\mathbf{x}; \beta) = \beta_0 + \beta_1 x_1 \quad (31)$$

pour obtenir les estimateurs ajustés sur le score de propension optimaux. Donc, les deux modèles étaient spécifiés correctement. Pour chaque échantillon, nous avons calculé quatre estimateurs de $\theta = N^{-1} \sum_{i=1}^N Y_i r_i$:

1. (ASP-EMV) : Estimateur ajusté sur le score de propension (3) avec $\hat{p}_i = p_i(\phi)$ et $\hat{\phi}$ étant l'estimateur du maximum de vraisemblance de ϕ .
2. (ASP-CAL) : Estimateur ajusté sur le score de propension (3) avec \hat{p}_i satisfaisant la contrainte de cailage (15) sur $(1, x_{z1})$.
3. (AUG) : Estimateur ajusté sur le score de propension augmenté (19).
4. (OPT) : Estimateur optimal (17).

Dans les estimateurs ajustés sur le score de propension augmenté, $\hat{\phi}$ a été calculé par la méthode du maximum de vraisemblance. Sous le modèle (30), l'estimateur du maximum de vraisemblance de $\phi = (\phi_0, \phi_1)'$ a été calculé en résolvant (6) avec $\mathbf{h}(\phi) = (1, x_{z1})'$. Le paramètre (β_0, β_1) par la méthode des moindres carrés ordinaires. En plus des estimateurs ponctuels, nous avons calculé les estimateurs de variance des estimateurs ajustés sur le score

où (b_0, b_1, γ_{h2}) est la limite de probabilité de $(\hat{b}_0, \hat{b}_1, \hat{\gamma}_{h2})$ ou

$$\eta_i(\phi, \beta) = p_i'(\phi) \mathbf{h}_i'(\phi, \beta) \gamma_{h*}^*$$

$$+ \frac{p_i'(\phi)}{r_i} \{ \gamma_i - p_i'(\phi) \mathbf{h}_i'(\phi, \beta) \gamma_{h*}^* \} \quad (24)$$

et l'effet de l'estimation de ϕ_0 dans $\hat{p}_i = p(\mathbf{x}_i; \hat{\phi})$ peut être ignoré sans risque.

Notons que, sous le modèle de réponse (4), $(\hat{\phi}, \hat{\lambda})$ dans (19) converge en probabilité vers $(\phi_0, 0)$, où ϕ_0 est le paramètre réel dans (4). Donc, le score de propension donne par le modèle augmente converge vers la probabilité de réponse réelle. Comme $\hat{\lambda}$ converge vers zéro en probabilité, le choix de β dans $\hat{m}_i = m(\mathbf{x}_i; \beta)$ ne joue aucun rôle dans l'absence asymptotique de biais de l'estimateur ajusté sur le score de propension. Les variances asymptotiques varient pour divers choix de β .

Sous le modèle de superpopulation (16), $\hat{b}_0 + \hat{b}_1 \mathbf{m}_i \rightarrow E(Y | \mathbf{x}_i)$ en probabilité. Donc, l'estimateur ajusté sur le score de propension optimal (19) est asymptotiquement équivalent à l'estimateur optimal (17). L'introduction de m_i dans l'équation de calage pour atteindre l'optimalité est proche de l'esprit de la méthode de calage sur un modèle proposée par Wu et Sitter (2001).

4. Estimation de la variance

Nous abordons maintenant l'estimation de la variance des estimateurs ajustés sur le score de propension sous le modèle de réponse supposé. Singh et Folsom (2000) et Kott (2006) ont discuté de l'estimation de la variance pour certains types d'estimateurs ajustés sur le score de propension. Kim et Kim (2007) ont discuté de l'estimation de la variance quand l'estimateur ajusté sur le score de propension est calculé par la méthode du maximum de vraisemblance.

Nous considérons l'estimation de la variance de l'estimateur ajusté sur le score de propension de la forme (3) où $\hat{p}_i = p(\cdot | \phi)$ est construite en vue de satisfaire (6) pour une certaine fonction $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi, \beta)$, où β est obtenu en utilisant le modèle de superpopulation positif. Soit β^* la limite de probabilité de β sous le modèle de réponse. Notons que β^* n'est pas nécessairement égal à β_0 dans (16), puisque nous ne supposons pas ici que le modèle de superpopulation positif est spécifié correctement.

Si nous utilisons de l'argument pour la linéarisation de Taylor (13) utilisés dans la preuve du théorème 1, l'estimateur ajusté sur le score de propension satisfait

$$\hat{\theta}^{\text{asp}} = \frac{1}{N} \sum_{i \in A} w_i \eta_i(\phi_0, \beta^*) + o_p(n^{-1/2}). \quad (23)$$

Pour obtenir l'estimateur de variance, nous supposons que l'estimateur de variance $V = N^{-2} \sum_{i \in A} \sum_{j \in A} \Omega_{ij}^{-1} \mathcal{G}_i \mathcal{G}_j'$ satisfait $V / V(\hat{\mathcal{G}}_{\text{HT}} | \mathcal{F}_N) = 1 + o_p(N)$ pour un certain Ω_{N-ij} . Nous supposons aussi que

$$|\Omega_{N-ij}^{-1}| = O(n^{-1}N). \quad (25)$$

Pour obtenir la variance totale, nous considérons le cadre inverse de Fay (1992), Shao et Sidel (1999) et Kim et Rao (2009). Dans ce cadre, la population finie est d'abord divisée en deux groupes, une population de répondants et une population de non-répondants. Connaissant la population de l'échantillon A est sélectionnée selon un plan d'échantillonnage probabiliste. Donc, la sélection de la population de répondants à partir de la population finie est traitée comme l'échantillonnage de première phase et la sélection de l'échantillon de répondants à partir de la population de non-répondants est traitée comme l'échantillonnage de deuxième phase dans le cadre inverse. La variance totale de $\hat{\eta}_{\text{HT}}$ peut s'écrire

$$V(\hat{\eta}_{\text{HT}} | \mathcal{F}_N) = V_1 + V_2 = E\{V(\hat{\eta}_{\text{HT}} | \mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\} + V\{E(\hat{\eta}_{\text{HT}} | \mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\}. \quad (26)$$

Dans (26), le terme de variance conditionnelle $V(\hat{\eta}_{\text{HT}} | \mathcal{F}_N, \mathcal{R}_N)$ peut être estimé par

où la dernière égalité découle du fait que y_i^* est conditionnellement indépendante de $E(Y|X_i) - p_i \mathbf{h}_i' \gamma_h^*$, en conditionnant sur \mathbf{x}_i . Puisque le dernier terme de (14) est non négatif, l'inégalité en (9) est établie. En outre, si $E(Y|X_i) = p_i \mathbf{h}_i' \alpha$ pour un certain α , (10) est vérifiée et $E(y_h^* | \mathbf{x}_i) = \alpha$, en vertu de la définition de y_h^* . Donc, $E(Y|X_i) - p_i \mathbf{h}_i' \gamma_h^* = -p_i \mathbf{h}_i' \gamma_h^* - E(y_h^* | \mathbf{x}_i) = o_p(1)$, ce qui implique que le dernier terme de (14) est négligeable.

Dans (9), V_i est la borne inférieure de la variance asymptotique des estimateurs ajustés sur le score de propension de la forme (3) satisfaisant (6). Tout estimateur ajusté sur le score de propension possédant la variance asymptotique V_i donnée en (9) est optimal puisqu'il atteint la borne inférieure de la variance asymptotique parmi la classe d'estimateurs ajustés sur le score de propension pour lesquels $\hat{\phi}$ satisfait (2). La variance asymptotique des estimateurs ajustés sur le score de propension optimaux de θ est égale à V_i en (9). L'estimateur ajusté sur le score de propension utilisant l'estimateur du maximum de vraisemblance de ϕ_0 n'atteint pas nécessairement la borne inférieure de la variance asymptotique.

La condition (10) donne un moyen de construire un estimateur ajusté sur le score de propension optimal. Premièrement, nous avons besoin d'une hypothèse pour $E(Y|X_i)$, qui est souvent appelée modèle de régression du résultat. Si le modèle de régression du résultat est un modèle de régression linéaire de la forme $E(Y|X_i) = \beta_0 + \beta_1' X_i$ un estimateur ajusté sur le score de propension optimal de θ peut être obtenu en résolvant

$$\sum_{i \in A} w_i \frac{p_i}{r_i} (\phi)(1, X_i) = \sum_{i \in A} w_i (1, X_i). \quad (15)$$

La condition (15) est intéressante, car elle dit que l'estimateur ajusté sur le score de propension appliqué à $y = a + \mathbf{b}' X_i$ mène à l'estimateur HT original. La condition (15) est appelée condition de calage dans le domaine des sondages. La condition de calage appliquée à \mathbf{x} utilise complètement l'information que celui-ci contient si la variable étudiée est bien approximée par une fonction linéaire de \mathbf{x} . La condition (15) a également été utilisée dans Neyo (2003) et dans Kott (2006) sous le modèle de régression linéaire. Si nous utilisons explicitement un modèle de régression pour $E(Y|X_i)$, il est possible de construire un estimateur qui possède la variance asymptotique (9) et n'est pas nécessairement un estimateur ajusté sur le score de propension. Par exemple, si nous supposons que

$$E(Y|X_i) = m(X_i; \beta_0) \quad (16)$$

$$\hat{\theta}_{\text{opt}}^* = \frac{1}{N} \sum_{i \in A} w_i \left[m(X_i; \beta) + \frac{p_i}{r_i} (\hat{\phi})(Y_i - m(X_i; \beta)) \right], \quad (17)$$

où $\hat{\theta}_{\text{opt}}^* = N^{-1} \sum_{i \in A} w_i \left[m(X_i; \beta_0) + \frac{p_i}{r_i} (\hat{\phi})(Y_i - m(X_i; \beta_0)) \right]$, et $V(\hat{\theta}_{\text{opt}}^*)$ est égal à V_i dans (9). Preuve. Définissons $\hat{\theta}_{\text{opt}}^*(\beta, \phi) = N^{-1} \sum_{i \in A} w_i [m(X_i; \beta) + \frac{p_i}{r_i} (\hat{\phi})(Y_i - m(X_i; \beta))]$. Notons que, dans (17), $\hat{\theta}_{\text{opt}}^*$ peut s'écrire $\hat{\theta}_{\text{opt}}^* = \hat{\theta}_{\text{opt}}^*(\beta, \phi)$. Puisque

$$\frac{\partial \hat{\theta}_{\text{opt}}^*}{\partial \beta} = \frac{1}{N} \sum_{i \in A} w_i \left\{ \frac{\partial m(X_i; \beta)}{\partial \beta} - \frac{p_i}{r_i} \frac{\partial (\hat{\phi})(Y_i - m(X_i; \beta))}{\partial \beta} \right\},$$

où $m(X_i; \beta) = \partial m(X_i; \beta) / \partial \beta$, et

$$\frac{\partial \hat{\theta}_{\text{opt}}^*}{\partial \phi} = \frac{1}{N} \sum_{i \in A} w_i p_i \frac{\partial (\hat{\phi})(Y_i - m(X_i; \beta))}{\partial \phi},$$

où $\hat{\theta}_{\text{opt}}^*(\beta, \phi) = \partial \{P_{-1}^{-1}(\phi)\} / \partial \phi$, nous avons $E[\partial \hat{\theta}_{\text{opt}}^*(\beta, \phi) / \partial \beta, \phi] = 0$ et la condition de Kandies (1982) est satisfaite. Donc,

$$\hat{\theta}_{\text{opt}}^*(\beta, \phi) = \hat{\theta}_{\text{opt}}^*(\beta_0, \phi) + o_p(n^{-1/2}) + o_p(n^{-1/2}) = \hat{\theta}_{\text{opt}}^*(\beta_0, \phi) + o_p(n^{-1/2})$$

et la variance de $\hat{\theta}_{\text{opt}}^*$ est égale à V_i , la borne inférieure de la variance asymptotique.

Théorème 2 Posons que les conditions du théorème 1 sont vérifiées. Supposons que β satisfait $\beta = \beta_0 + O_p(n^{-1/2})$. Supposons que $m(X; \beta)$ possède des dérivées partielles d'ordre un continues dans un ensemble ouvert contenant β_0 . Sous la distribution conjointe du mécanisme d'échantillonnage, du mécanisme de réponse et du modèle de superpopulation (16), l'estimateur $\hat{\theta}_{\text{opt}}^*$ en (17) satisfait

atteint la borne inférieure en (9).

Le théorème qui suit montre que l'estimateur optimal (17) mateur convergent à la vitesse $\sim n$ de ϕ_0 calculé par (6), dans le modèle de superpopulation (16) et $\hat{\phi}$ est un estimateur convergent à la vitesse $\sim n$ de β_0

Dans (9), l'égalité tient quand $\hat{\Phi}_h$ satisfait

$$(10) \quad \sum_{i \in I} w_i \left\{ \frac{p(\mathbf{x}_i; \hat{\Phi}_h)}{r_i} - 1 \right\} E(Y | \mathbf{x}_i) = 0,$$

où $E(Y | \mathbf{x}_i)$ est l'espérance conditionnelle sous le modèle de superpopulation.

Preuve. Étant donné $p_i(\Phi) = p(\mathbf{x}_i; \Phi)$ et $\mathbf{h}_i(\Phi) = \mathbf{h}(\mathbf{x}_i; \Phi)$, définissons

$$\theta(\Phi, \gamma) =$$

$$N^{-1} \sum_{i \in I} w_i \left[p_i(\Phi) \mathbf{h}_i'(\Phi) \mathbf{h}_i(\Phi) \gamma + \frac{d_i(\Phi)}{r_i} \{ \gamma_i - p_i(\Phi) \} \gamma \right].$$

Puisque $\hat{\Phi}_h$ satisfait (6), nous avons $\hat{\theta}(\hat{\Phi}_h, \gamma) = \hat{\theta}(\Phi_0, \gamma)$ pour tout choix de γ . Nous voulons maintenant trouver un choix particulier de γ , disons γ^* , tel que

$$(11) \quad \hat{\theta}(\hat{\Phi}_h, \gamma^*) = \hat{\theta}(\Phi_0, \gamma^*) + o_p(n^{-1/2}).$$

Comme $\hat{\Phi}_h$ converge en probabilité vers Φ_0 , l'équivalence asymptotique (11) est vérifiée si

$$(12) \quad E \left\{ \frac{\partial \hat{\theta}}{\partial \Phi} (\Phi, \gamma^*) \mid \Phi = \Phi_0 \right\} = \mathbf{0},$$

en utilisant la théorie de Randles (1982). La condition (12) est vérifiée si $\gamma^* = \gamma_h^*$ où γ_h^* est défini en (8). Donc, (11)

se réduit à

$$\hat{\theta}_{\text{ASP}, h} = \frac{1}{N} \sum_{i \in I} w_i \left\{ p_i' \mathbf{h}_i' \gamma_h^* + \frac{d_i'}{r_i} (\gamma_i - p_i') \mathbf{h}_i' \gamma_h^* \right\} + o_p(n^{-1/2}),$$

(13)

ce qui prouve (7). La variance de $\hat{\theta}_{\text{ASP}, h}$ peut être calculée par

$$V(\hat{\theta}_{\text{ASP}, h})$$

$$= V(\hat{\theta}^{\text{HT}}) + \frac{N}{1} \sum_{i \in I} w_i^2 \left(\frac{d_i'}{1} - 1 \right) (Y_i - p_i') \left\{ \frac{d_i'}{1} - 1 \right\} E(Y | \mathbf{x}_i) + E(Y | \mathbf{x}_i) - p_i' \mathbf{h}_i' \gamma_h^* \left\{ \frac{d_i'}{1} - 1 \right\}^2$$

$$+ \frac{N}{1} \sum_{i \in I} w_i^2 \left(\frac{d_i'}{1} - 1 \right) E \left\{ E(Y | \mathbf{x}_i) - p_i' \mathbf{h}_i' \gamma_h^* \right\}^2. \quad (14)$$

d'échantillonnage sont bornés uniformément. Autrement dit, $K_1 < N^{-1} n_{w_i} < K_2$ pour tout i uniformément dans n , où K_1 et K_2 sont des constantes données. En outre, nous supposons qu'il existe les conditions de régularité suivantes :

[C1] Le mécanisme de réponse satisfait (4), où $p(\mathbf{x}; \Phi)$ est continue en Φ avec les dérivées première et seconde continues dans un ensemble ouvert contenant Φ_0 . $\text{Cov}(r_i, r_j | \mathbf{x}) = 0$ pour $i \neq j$. En outre, $p(\mathbf{x}_i; \Phi) > c$ pour tout i pour une certaine constante donnée $c > 0$.

[C2] La solution de (6) existe et est unique presque partout. Dans (6), la fonction $\mathbf{h}_i(\Phi) = \mathbf{h}(\mathbf{x}_i; \Phi)$ possède un moment de quatrième ordre borné. En outre, la dérivée partielle $\partial \{ \mathbf{h}_i(\Phi) \} / \partial \Phi$ est non singulière pour tout n .

[C3] Dans (6), la fonction d'estimation $\hat{\mathbf{U}}_h(\Phi)$ converge en probabilité vers $\mathbf{U}_h(\Phi) = \sum_{i=1}^N \{ r_i - p_i(\Phi) \} \mathbf{h}_i(\Phi)$ uniformément en Φ . En outre, la dérivée partielle $\partial \{ \mathbf{U}_h(\Phi) \} / \partial \Phi$ converge en probabilité vers $\partial \{ \mathbf{h}_i(\Phi) \} / \partial \Phi$ uniformément en Φ . La solution Φ_N^d de $\mathbf{U}_h^d(\Phi) = \mathbf{0}$ satisfait $N^{1/2}(\Phi_N^d - \Phi_0) = O_p(1)$ sous le mécanisme de réponse.

La condition [C1] énonce les conditions de régularité pour le mécanisme de réponse. La condition [C2] est la condition de régularité pour la solution $\hat{\Phi}_h$ de (6). Dans la condition [C3], certaines conditions de régularité sont imposées à la fonction d'estimation $\hat{\mathbf{U}}_h(\Phi)$ proprement dite. Par [C2] et [C3], nous pouvons établir la convergence à la vitesse $n^{-1/2}$ de $\hat{\Phi}_h$ en (2).

Maintenant, le théorème qui suit traite de certaines propriétés asymptotiques de l'estimateur ajusté sur le score de propension $\hat{\theta}_{\text{ASP}, h}$.

Théorème 1 Si les conditions [C1] à [C3] sont vérifiées, sous la distribution conjointe du mécanisme d'échantillonnage et du mécanisme de réponse, l'estimateur ajusté sur le score de propension $\hat{\theta}_{\text{ASP}, h}$ satisfait

$$(7) \quad n(\hat{\theta}_{\text{ASP}, h} - \theta_{\text{ASP}, h}) = o_p(1),$$

$$\hat{\theta}_{\text{ASP}, h} = \frac{1}{N} \sum_{i \in I} w_i \left\{ p_i' \mathbf{h}_i' \gamma_h^* + \frac{d_i'}{r_i} (\gamma_i - p_i') \mathbf{h}_i' \gamma_h^* \right\} + o_p(n^{-1/2}). \quad (8)$$

où $\gamma_h^* = (\sum_{i=1}^N \{ r_i - p_i(\Phi) \} \mathbf{h}_i(\Phi))^{-1} (\sum_{i=1}^N \{ r_i - p_i(\Phi) \} \mathbf{h}_i(\Phi) \gamma_h^*) / \partial \Phi$, et $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i; \Phi_0)$. De plus, si la population finie est un échantillon aléatoire tiré d'un modèle de superpopulation, alors

$$V(\hat{\theta}_{\text{ASP}, h}) \equiv V(\hat{\theta}^{\text{HT}}) + \frac{N}{1} \sum_{i \in I} w_i^2 \left(\frac{d_i'}{1} - 1 \right) E \left\{ E(Y | \mathbf{x}_i) - p_i' \mathbf{h}_i' \gamma_h^* \right\}^2. \quad (9)$$

Kim et Riddles : Théorie concernant les estimateurs ajustés sur le score de propension dans les sondages

de régression logistique pour estimer la probabilité de

le score de propension à une enquête en ligne auprès d'un panel de volontaires. Durrant et Skinner (2006) ont utilisé l'approche d'ajustement sur le score de propension pour traiter l'erreur de mesure.

Malgré la popularité des estimateurs ajustés sur le score de propension, peu d'attention a été accordée à leurs propriétés asymptotiques dans la littérature sur les sondages. Kim et Kim (2007) ont utilisé un développement en série de Taylor pour obtenir la moyenne et la variance asymptotique des estimateurs ajustés sur le score de propension et discuté de l'estimation de leur variance. Da Silva et Opsomer (2006), ainsi que Da Silva et Opsomer (2009) ont considéré des méthodes non paramétriques pour obtenir des estimateurs ajustés sur le score de propension.

Dans le présent article, nous discutons des estimateurs ajustés sur le score de propension optimaux appartenant à la classe d'estimateurs de la forme (3) qui utilisent un estimateur $\hat{\phi}$ convergeant à la vitesse \sqrt{n} . Ces estimateurs sont asymptotiquement sans biais pour θ à la recherche

ajustés sur le score de propension est l'un des principaux sujets sur lesquels porte le présent article.

À la section 2, nous présentons les résultats principaux. À la section 3, nous proposons un estimateur ajusté sur le score de proportion optimal utilisant un modèle de score de propension augmenté. À la section 4, nous discutons de l'estimation de la variance de l'estimateur proposé. À la section 5, nous présentons les résultats de deux études par simulation et à la section 6, nous formulons nos conclusions.

2. Résultats principaux

A la présente section, nous discutons de certaines propriétés asymptotiques des estimateurs ajustés sur le score de propension. Nous supposons que le mécanisme de réponse ne dépend pas de y , donc que

$$(4) \quad \Pr(\phi^0; \mathbf{x}) d = \Pr(y = 1 | \mathbf{x})$$

pour un certain vecteur ϕ_0 inconnu. La première égalité implique que les données manquent au hasard (MAR pour *missing-at-random*), car nous observons toujours x dans l'échantillon. Notons que la condition MAR fait partie des hypothèses du modèle de population. Dans la seconde égalité, nous supposons en outre que le mécanisme de réponse est connu jusqu'à un paramètre ϕ_0 inconnu. Le mécanisme de réponse est légèrement différent de celui

$$\Pr(y, I = 1 | \mathbf{x}, I = 1) = \Pr(I = 1 | \mathbf{x}, I = 1) d(\mathbf{x}; \phi_0^y). \quad (5)$$

tion, comme il est discuté à la section 1. Nous considérons une classe d'estimateurs convergeant à la vitesse $n^{-\phi}$ dans (4). En particulier, nous considérons une classe d'estimateurs qui peuvent s'écrire comme une solution de

$$(9) \quad \theta = (\phi)^t \mathbf{u} \{ (\phi)^t d - \mathcal{A} \}^t \mathcal{M} \sum_{V \ni !} \equiv (\phi)^t \mathbf{u}$$

où $d^i d^j = (\phi)^i(\phi)^j d^j d^i$ pour une certaine fonction ϕ , $\phi(\phi)^i(\phi)^j = \mathbf{h}(\mathbf{x})^i \mathbf{x}^j$, une fonction lisse de \mathbf{x}^i et du paramètre ϕ . Donc, la solution de (6) peut s'écrire comme étant $\phi^{n/2}(\phi)$, qui dépend du choix de $\mathbf{h}(\phi)$. Toute solution $\phi^{n/2}(\phi)$ de (6) est

convergente pour Φ dans (4) parce que $\{E_i(\Phi)U_i\}^{\mathcal{N}} = \{E_i(\Phi)U_i\}^{\mathcal{N}'} = \emptyset$ et $\{E_i(\Phi)U_i\}^{\mathcal{N} \cup \mathcal{N}'} = \{E_i(\Phi)U_i\}^{\mathcal{N}} \cup \{E_i(\Phi)U_i\}^{\mathcal{N}'}$.
 (5) Φ est convergente pour Φ dans (5) et l'estimateur ajusté

sur le score de propension résultant est convergent unique-
ment si le plan d'échantillonnage est non informatif. Les
estimations ajustées sur le score de propension obtenu à partir
de (6) en utilisant les poids d'échantillonnage sont conver-
gents que le plan d'échantillonnage soit ou ne soit pas non-
informatif. Selon Chamberlain (1987), tout estimateur con-
vergeant à la vitesse \sqrt{n} de ϕ_0 dans (4) peut s'écrire sous
la forme d'une solution de (6). Donc, le choix de $h_1(\phi)$
dans (6) détermine l'efficacité de l'estimateur ajusté sur le

Soit $\hat{\theta}_{\text{Asp},H}^{\text{Asp}}$ l'estimateur ajusté sur le score de propension donné par (3) en utilisant $\hat{p}_i = p^*(\Phi_i)$ où Φ_i est la solution de (6). Pour discuter des propriétés asymptotiques de $\hat{\theta}_{\text{Asp},H}^{\text{Asp}}$ supposons que nous ayons une suite de populations finies et d'échantillons, comme dans Isaki et Fuller (1982), telle que $\sum_{i \in \mathcal{N}} w_i^* \mathbf{1}_i^T = \sum_{i \in \mathcal{N}} \mathbf{1}_i^T = O(n^{1-1/2})$ caractéristique de la population $\mathbf{1}_i^T$ avec les moments d'ordre quatre bornés. Nous supposons aussi que les poids

Théorie concernant les estimateurs ajustés sur le score de propension dans les sondages

Jae Kwang Kim et Minsun Kim Riddles¹

Résumé

La méthode d'ajustement sur le score de propension est souvent adoptée pour traiter le biais de sélection dans les sondages, y compris la non-réponse totale et le sous-dénombrement. Le score de propension est calculé en se servant de variables auxiliaires observées dans tout l'échantillon. Nous discutons de certaines propriétés asymptotiques des estimateurs ajustés sur le score de propension et dérivons des estimateurs optimaux fondés sur un modèle de régression pour la population finie. Un estimateur ajusté sur le score de propension optimal peut être réalisé en se servant d'un modèle de score de propension augmenté. Nous discutons de l'estimation de la variance et présentons les résultats de deux études par simulation.

Mots clés : Catage ; données manquantes ; non-réponse ; pondération.

1. Introduction

Considérons une population finie de taille N , où N est connu. Pour chaque unité i , y_i est la variable étudiée et \mathbf{x}_i est le vecteur de dimension q de variables auxiliaires. Le paramètre d'intérêt est la moyenne de population finie de la variable étudiée, $\theta = N^{-1} \sum_{i=1}^N y_i$. Supposons que la population finie $\mathcal{F}_N = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$ est un échantillon aléatoire de taille N tiré d'une loi de super-

population $F(\mathbf{x}, y)$. Supposons qu'un échantillon de taille n est tiré de la population finie selon un plan d'échantillonnage probabiliste. Soit $w_i = \pi_i^{-1}$ le poids d'échantillonnage, où π_i est la probabilité d'inclusion de premier ordre de l'unité i obtenue d'après le plan d'échantillonnage probabiliste. Sous réponse complète, la moyenne de population finie peut être estimée par l'estimateur d'Horvitz-Thompson (HT), $\hat{\theta}_{HT} = N^{-1} \sum_{i \in A} w_i y_i$, où A est l'ensemble d'indices qui figurent dans l'échantillon.

En présence de données manquantes, l'estimateur HT $\hat{\theta}_{HT}$ ne peut pas être calculé. Soit r la variable indicatrice de réponse, qui prend la valeur 1 si y est observée et la valeur 0 autrement. Conceptuellement, comme on ont discuté Fay (1992), Shao et Steel (1999), ainsi que Kim et Rao (2009), l'indicateur de réponse peut être étendu à la population entière sous la forme $\mathcal{R}_N = \{r_1, r_2, \dots, r_N\}$, où r_i est une réalisation de la variable aléatoire r_i . L'estimateur sous cas complets (CC) $\hat{\theta}_{CC} = \sum_{i \in A} w_i r_i y_i / \sum_{i \in A} w_i r_i$ converge alors en probabilité vers $E(Y | r = 1)$. À moins que le mécanisme de réponse soit tel que les réponses manquent entièrement au hasard en ce sens que $E(Y | r = 1) = E(Y)$, l'estimateur CC présente un biais. Pour corriger ce biais, si la probabilité de réponse

$$p(\mathbf{x}, y) = \Pr(r = 1 | \mathbf{x}, y) \quad (1)$$

est connue, on peut utiliser l'estimateur CC pondéré $\hat{\theta}_{CCP} = N^{-1} \sum_{i \in A} w_i r_i y_i / p(\mathbf{x}_i, y_i)$ pour estimer θ . Il convient de souligner que $\hat{\theta}_{CCP}$ est sans biais parce que $E\{\sum_{i \in A} w_i r_i y_i / p(\mathbf{x}_i, y_i) | \mathcal{F}_N\} = E\{\sum_{i=1}^N y_i\} = \theta$. Si la probabilité de réponse (1) est inconnue, on peut postuler pour cette dernière un modèle paramétrique $p(\mathbf{x}, y; \phi)$ indexé par $\phi \in \Omega$ tel que $p(\mathbf{x}, y) = p(\mathbf{x}, y; \phi_0)$ pour un certain $\phi_0 \in \Omega$. Nous supposons qu'il existe un estimateur convergent à la vitesse $\sim n^{-1/2}$ de ϕ_0 tel que

$$n(\hat{\phi} - \phi_0) = O_p(1), \quad (2)$$

où $\mathcal{G}_n = O_p(1)$ indique que \mathcal{G}_n est borné en probabilité. En utilisant $\hat{\phi}$, nous pouvons obtenir la probabilité de réponse estimée par $\hat{p}_i = p(\mathbf{x}_i, y_i; \hat{\phi})$, qui est souvent appelée score de propension (Rosenbaum et Rubin 1983). L'estimateur ajusté sur le score de propension (ASP) peut être construit comme

$$\hat{\theta}_{ASP} = \frac{1}{N} \sum_{i \in A} w_i \frac{y_i}{\hat{p}_i} \quad (3)$$

L'estimateur ajusté sur le score de propension (3) est d'usage très répandu. De nombreux programmes d'enquête l'utilisent pour réduire le biais de non-réponse (Fuller, Loughlin et Baker, 1994; Rizzo, Kalton et Brick, 1996). Rosenbaum et Rubin (1983) et Rosenbaum (1987) ont proposé d'utiliser l'approche d'ajustement sur le score de propension pour estimer les effets du traitement dans les études observationnelles. Little (1988) a passé en revue les méthodes d'ajustement sur le score de propension pour le traitement de la non-réponse totale dans les sondages. Duncan et Stasny (2001) ont utilisé l'approche d'ajustement sur le score de propension pour contrôler le biais de couverture

- Diggle, P., et Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 43, 49-93.
- Duan, N., et Li, K.C. (1991). Sliced regression: A link-free regression method. *The Annals of Statistics*, 19, 505-530.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.
- Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1, 1-17.
- Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.
- Little, R.J., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, deuxième édition. New York : John Wiley & Sons, Inc.
- National Science Foundation, Division of Science Resources Statistics (2010). *Research and Development in Industry: 2005. Detailed Statistical Tables*. Disponible au <http://www.nsf.gov/statistics/nsf10319/>.
- Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92, 1320-1329.
- R Development Core Team (2009). *A language and environment for statistical computing, R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0.
- Robins, J.M., et Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122-129.
- Shao, J., et Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.
- Troxel, A.B., Harrington, D.P., et Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, 47, 425-438.
- Troxel, A.B., Lipsitz, S.R., et Harrington, D.P. (1998). Marginal models for the analysis of longitudinal measurements with non-ignorable non-monotone missing data. *Biometrika*, 85, 661-672.
- Vansteelandt, S., Rotnitzky, A., et Robins, J.M. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94, 841-860.
- Xu, J. (2007). Methods for intermittent missing responses in longitudinal data. Thèse de doctorat, Department of Statistics, University of Wisconsin-Madison.
- Xu, J., Shao, J., Palla, M., et Wang, L. (2008). Imputation pour les enquêtes longitudinales. *Techniques d'enquête*, 34, 2, 169-179.

Un exemple dans lequel (4) n'est pas vérifiée. Pour montrer que (4) n'est pas vérifiée en général, il nous suffit de donner un contre-exemple. Considérons $T = 3$. Posons que (Y_1, Y_2, Y_3) suivent conjointement une loi normale caractérisée par $E(Y_i) = 0$, $\text{var}(Y_i) = 1$, $i = 1, 2, 3$, $\text{cov}(Y_1, Y_2) = \text{cov}(Y_1, Y_3) = \rho^2$, et $\text{cov}(Y_2, Y_3) = \rho^2$, où $\rho \neq 0$ est un paramètre. Supposons que Y_1 est toujours observée et que $P(\delta_i = 0 | Y^{(-i)}) = \Phi(a_{i-1} + b_{i-1}Y^{(-i)})$, $i = 2, 3$, où a_i et b_i sont des paramètres, et Φ est la fonction de répartition de la loi normale centrée réduite. Alors $E(Y_3 | Y^2, Y_1) = p_{Y_3}, E(Y_2 | Y_1) = p_{Y_1}$, et $E(Y_3 | Y_1) = p^2 Y_1$. Notons que

$$E(Y_3 | Y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) = E(Y_3 | Y_1, \delta_3 = 0, \delta_2 = 1)$$

$$= \int E(Y_3 | Y_1, \delta_3 = 0) dY_3 = \int \int \int Y_3 L(Y_3 | Y_1, Y_2, \delta_3 = 0) L(Y_2 | Y_1, \delta_3 = 0) dY_2 dY_3$$

$$= \left(\int \int Y_3 L(Y_3 | Y_2, \delta_3 = 0) dY_3 \right) L(Y_2 | Y_1, \delta_3 = 0) dY_2$$

$$= \int p dY_2 L(Y_2 | Y_1, \delta_3 = 0) dY_2$$

$$= \int p Y_2 L(Y_2 | Y_1, \delta_3 = 0) dY_2 = 0$$

$$= \int p \int Y_2 \Phi(a_2 + b_2 Y_2) L(Y_2 | Y_1) dY_2 = \int p \int Y_2 \Phi(a_2 + b_2 Y_2) L(Y_2 | Y_1) dY_2$$

où la première égalité est vérifiée parce que Y_1 est toujours observée, la deuxième égalité est vérifiée parce que, sous (1), δ_2 et Y_3 sont indépendants sachant Y_1 . Le dénominateur de l'expression précédente est égal à

$$h(Y_1) = \Phi \left(\frac{a_2 + b_2 p Y_1}{a_2^2 + b_2^2 p^2 Y_1^2} \right) \left(1 - p^2 \right)$$

En intégrant par parties, nous obtenons que

$$E(Y_1) = \int (Y_2 - p Y_1) \Phi(a_2 + b_2 Y_2) L(Y_2 | Y_1) dY_2$$

$$= b_2 (1 - p^2) \int \Phi(a_2 + b_2 Y_2) L(Y_2 | Y_1) dY_2$$

$$= \frac{b_2^2 (1 - p^2)}{2} \int \exp \left\{ - \frac{2}{(a_2^2 + b_2^2 p^2 Y_2^2)} \right\} dY_2 - \frac{2(1 - p^2)}{(Y_2 - p Y_1)^2} \left\{ \exp \left[\frac{2(1 - p^2)}{(a_2^2 + b_2^2 p^2 Y_2^2)} \right] \right\}$$

D'où,

$$E(Y_3 | Y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) = p^2 Y_1 + p \frac{E(Y_1)}{E(Y_1)} \quad (10)$$

Cependant, $E(Y_3 | Y_1, \delta_1 = \delta_2 = 1) = E(Y_3 | Y_1, \delta_1 = 1) = E(Y_3 | Y_1) = p^2 Y_1$. Cela montre que (4) n'est pas vérifiée dans ce cas particulier.

Preuve de (8). En utilisant la notation de la preuve de (2) et de (3), et en représentant le vecteur $(Y_1^{(-1)}, \dots, Y_{t-1}^{(-1)}, Y_t^{(-1)}, \dots, Y_{t-1}^{(-1)})$ de dimension $(t - 2)$ par $\mathbf{n}^{t,r}$, nous obtenons

$$L(\delta_{t+1} = 1 | Y^{t,r}, \Delta_r = 1, \delta_t = 0)$$

$$= \int L(\delta_{t+1} = 1 | Y^{t,r}, \mathbf{n}^{t,r}, \Delta_r = 1, \delta_t = 0) d\mathbf{n}^{t,r}$$

$$= \int L(\delta_{t+1} = 1 | Y_1, \dots, Y^r, \Delta_r = 1) dY^r$$

$$= \int L(\mathbf{n}^{t,r} | Y^{t,r}, \Delta_r = 1, \delta_t = 0) d\mathbf{n}^{t,r}$$

$$= \int L(\delta_{t+1} = 1 | Z^r, \Delta_r = 1) dZ^r$$

$$= L(\delta_{t+1} = 1 | Z^r, \Delta_r = 1)$$

$$= \int L(\mathbf{n}^{t,r} | Y^{t,r}, \Delta_r = 1, \delta_t = 0) d\mathbf{n}^{t,r}$$

$$= L(\delta_{t+1} = 1 | Z^r, \Delta_r = 1)$$

où la deuxième égalité découle de l'hypothèse (1) et du fait qu'il existe une fonction injective entre $(Z^r, \mathbf{n}^{t,r})$ et $(Y_1^{(-1)}, \dots, Y_{t-1}^{(-1)})$, et la troisième égalité découle de l'hypothèse (7). De même, $L(\delta_{t+1} = 1 | Z^r, \Delta_r = 1, \delta_t = 0) = L(\delta_{t+1} = 1 | Z^r, \Delta_r = 1)$ et, donc, $L(\delta_{t+1} = 1 | Y_1^{(-1)}, \dots, Y_{t-1}^{(-1)}, \Delta_r = 1, \delta_t = 0) = L(\delta_{t+1} = 1 | Z^r, \Delta_r = 1, \delta_t = 0)$. Alors,

$$L(Y_t | Z^r, \Delta_{t+1} = 1, \delta_t = 0)$$

$$= \frac{L(Y_t | Z^r, \Delta_{t+1} = 1, \delta_t = 0)}{L(Z^r, \Delta_{t+1} = 1, \delta_t = 0)}$$

$$= \frac{L(\delta_{t+1} = 1 | Y_t, Z^r, \Delta_r = 1, \delta_t = 0)}{L(\delta_{t+1} = 1 | Z^r, \Delta_r = 1, \delta_t = 0)} L(Z^r, \Delta_r = 1, \delta_t = 0)$$

$$= L(Y_t | Z^r, \Delta_r = 1, \delta_t = 0)$$

De même, $L(Y_t | Z^r, \Delta_r = 1, \delta_{t+1} = 0, \delta_t = 0) = L(Y_t | Z^r, \Delta_r = 1, \delta_t = 0)$. D'où, $L(Y_t | Z^r, \Delta_r = 1, \delta_{t+1} = 0, \delta_t = 0) = L(Y_t | Z^r, \Delta_{t+1} = 1, \delta_t = 0)$ et le résultat (8) s'ensuit.

Bibliographie

Bond, D. (1994). An evaluation of imputation methods for the Survey of Industrial Research and Development. *U.S. Bureau of the Census, Economic Statistical Methods and Programming Division Report Series*, 9404. Washington, DC.

Cheng, P.F. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89, 81-87.

covariables :

thèse (1) peut être modifiée comme il suit afin d'inclure les une covariable observée x_i sans valeur manquante. L'hypothèse (1) peut être modifiée comme il suit afin d'inclure les situation où chaque unité échantillonnée possède au temps t Les résultats de la section 2 peuvent être étendus à la linéaire proposée.

mais elles sont bonnes sous l'imputation par la régression diocres dans les conditions des simulations de la section 3.1, sur la population de la SIRD, leurs propriétés sont meilleures que les conditions des simulations de la section 3.1, sont en général asymptomatiquement biaisées. Même s'ils tation par la régression linéaire (8) et (9) décrites à la section Les estimateurs qui s'appuient sur les méthodes d'imputation méthodes proposées.

estimateurs fondés sur la censure à ceux fondés sur les Dans les applications, il pourrait être utile de comparer les données recueillies pour un plus grand nombre d'années. aboutir à des estimateurs inefficaces si l'on considère les lles pour quatre années seulement et la censure pourrait Cependant, ces résultats sont fondés sur les données recueillies méthode d'imputation par la régression linéaire proposée. 3.2 et 3.3, la censure donne des résultats comparables à la grand. Pour l'analyse des données de la SIRD aux sections sinon, elle peut être très inefficace, surtout quand T est résultats quand le nombre de données écartées est faible ; valeur manquante chez ce dernier, peut donner de bons données observées auprès d'un sujet après la première La méthode de censure, qui consiste à écarter toutes les que les méthodes fondées sur la régression à noyau.

est plus stable et, par conséquent, peut demeurer meilleure fondée sur la régression linéaire est asymptotiquement biaisée quand la relation linéaire n'est pas vérifiée, mais elle tions pour chaque schéma de non-réponse. L'imputation ment valide, mais elle requiert un grand nombre d'observations fondée sur la régression à noyau est asymptotiquement à noyau avec indice unidimensionnel. La méthode d'imputation la régression linéaire, la régression à noyau, et la régression à noyau avec indice unidimensionnel. La méthode d'imputation fondée sur la régression à noyau est asymptotiquement biaisée quand la relation linéaire n'est pas vérifiée, mais elle que les méthodes fondées sur la régression à noyau.

4. Conclusion

résultat (5) s'ensuit.

$\delta_{j-1} = 0, \delta_j = 0) = L(y_j | y_{j-1}, \Delta_j = 1, \delta_{j-1} = 1, \delta_j = 0)$ et le nous pouvons montrer que $L(y_j | y_{j-1}, \Delta_j = 1, \delta_{j-1} = 1, \delta_j = 1, \delta_{j+1} = 1, \delta_{j+2} = 0) = L(y_j | y_{j-1}, \Delta_j = 1, \delta_{j-1} = 1, \delta_j = 1, \delta_{j+1} = 1, \delta_{j+2} = 0)$ en vertu de (1). De même, à $L(y_j | y_{j-1}, \Delta_j = 1, \delta_{j-1} = 1, \delta_j = 0, \delta_{j+1} = 1, \delta_{j+2} = 0)$ et en posant que $\Delta_j = 1$ est l'indicateur de preuve de (2) et en utilisant la même notation que dans la *Preuve de (5)*. En utilisant la même notation que dans la coulent de l'hypothèse (1).

$\dots = L(y_j | y_{j-1}, \Delta_j = 1, \delta_{j-1} = 1, \delta_j = 0, \delta_{j+1} = 1, \delta_{j+2} = 0)$ et en posant que $\Delta_j = 1$ est l'indicateur de preuve de (2) et en utilisant la même notation que dans la coulent de l'hypothèse (1).

Annexe

David L. Kinyon, tous deux du U.S. Census Bureau, ainsi que deux examinateurs et le rédacteur associé de leurs nombreux commentaires constructifs au sujet de l'article. L'étude a été financée en partie par une subvention de la NSF. Le présent article est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche en cours et de favoriser la discussion. Les opinions exprimées sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau.

Remerciements

Nous remercions Katherine Jenny Thompson et David L. Kinyon, tous deux du U.S. Census Bureau, ainsi que deux examinateurs et le rédacteur associé de leurs nombreux commentaires constructifs au sujet de l'article. L'étude a été financée en partie par une subvention de la NSF. Le présent article est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche en cours et de favoriser la discussion. Les opinions exprimées sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau.

$P(\delta_j = 1 | y, X, \delta_{j-1}, \dots, \delta_{j+1}, \dots, \delta_T) = P(\delta_j = 1 | y_j, X_j, \delta_{j-1}, \dots, \delta_{j+1}, \dots, \delta_T) = P(\delta_j = 1 | y_j, X_j, \delta_{j-1}, \dots, \delta_{j+1}, \dots, \delta_T)$

1. Contrairement aux observations dans les conditions de simulation d'une population normale, le total estimé fondé sur la censure et la régression linéaire possède un écart-type comparable à celui obtenu par les méthodes d'imputation proposées. Il en est ainsi parce que le nombre de points de données écartés par censure est faible dans le cas qui nous occupe. Les probabilités d'un schéma de réponse intermittente sont de 17 % et 19 % pour les classes des strates à tirage complet et à tirage partiel, respectivement. Dans la simulation de la population normale, ces probabilités s'approchaient de 50 % comme le montre le tableau 2.

2. Toutes les méthodes d'imputation proposées donnent des résultats relativement semblables. Comme nous

Tableau 5
Résultats de simulation pour les estimations du total (en milliers) pour la population fondée sur la SIRD

Méthode	Quantité	Strates à tirage complet				Strates à tirage partiel			
		$t = 2$	$t = 3$	$t = 4$	$t = 2$	$t = 3$	$t = 4$	$t = 2$	$t = 3$
Données complètes	Biais relatif	0 %	0,1 %	0,1 %	0,2 %	0,0 %	0,4 %	0,4 %	0,4 %
	$E-T$	15 541	16 045	16 947	18 484	203	184	224	224
	$E-T_{boot}$	15 654	15 994	16 941	186	201	218	218	218
	Couverture de l'IC	94,0 %	94,0 %	94,3 %	94,3 %	93,7 %	93,9 %	93,9 %	93,9 %
Répondants seulement avec poids corrigés	Biais relatif	5 %	6,3 %	11,6 %	-1,1 %	1,1 %	-2,7 %	-2,7 %	-2,7 %
	$E-T$	16 870	17 858	20 032	191	220	244	244	244
	$E-T_{boot}$	17 917	17 915	20 048	192	219	234	234	234
	Couverture de l'IC	94,8 %	94,8 %	87,3 %	93,2 %	94,5 %	89,8 %	89,8 %	89,8 %
Censure et imputation par la régression linéaire	Biais relatif	0 %	0,4 %	0,5 %	0,4 %	0,1 %	-0,4 %	-0,4 %	-0,4 %
	$E-T$	15 582	16 272	17 247	191	214	238	238	238
	$E-T_{boot}$	15 654	16 145	17 195	194	214	236	236	236
	Couverture de l'IC	93,8 %	93,5 %	94,2 %	94,8 %	94,0 %	93,7 %	93,7 %	93,7 %
Imputation par la régression linéaire proposée	Biais relatif	0 %	0,2 %	0,0 %	0,4 %	0,0 %	-0,5 %	-0,5 %	-0,5 %
	$E-T$	15 582	16 130	16 955	191	206	229	229	229
	$E-T_{boot}$	15 654	16 072	16 964	194	206	224	224	224
	Couverture de l'IC	93,8 %	93,5 %	94,2 %	94,8 %	94,0 %	93,7 %	93,7 %	93,7 %
Imputation par la régression à noyau avec indice unidimensionnel proposée	Biais relatif	0 %	0,2 %	-0,1 %	0,4 %	-0,3 %	-0,3 %	-0,3 %	-0,3 %
	$E-T$	15 582	16 130	16 957	191	205	227	227	227
	$E-T_{boot}$	15 654	16 072	16 965	194	204	220	220	220
	Couverture de l'IC	93,8 %	93,5 %	94,3 %	94,8 %	93,4 %	93,1 %	93,1 %	93,1 %
Imputation par la régression à noyau avec dépendance à l'égard de la dernière valeur	Biais relatif	0 %	0,1 %	-0,3 %	0,0 %	-0,7 %	-0,7 %	-0,7 %	-0,7 %
	$E-T$	15 565	16 019	16 990	184	204	242	242	242
	$E-T_{boot}$	15 635	16 003	16 983	187	202	230	230	230
	Couverture de l'IC	93,8 %	93,7 %	94,0 %	93,9 %	92,7 %	91,1 %	91,1 %	91,1 %
Imputation par la régression linéaire avec valeurs imputées antérieures	Biais relatif	0 %	0,2 %	0,0 %	0,4 %	0,6 %	-0,6 %	-0,6 %	-0,6 %
	$E-T$	15 582	16 120	16 952	191	210	231	231	231
	$E-T_{boot}$	15 654	16 065	16 954	194	210	225	225	225
	Couverture de l'IC	93,8 %	93,6 %	94,3 %	94,8 %	93,8 %	92,8 %	92,8 %	92,8 %
Imputation par la régression linéaire fondée sur les données observées courantes et antérieures	Biais relatif	0 %	0,2 %	0,0 %	0,4 %	0,6 %	-0,6 %	-0,6 %	-0,6 %
	$E-T$	15 582	16 117	16 945	191	213	241	241	241
	$E-T_{boot}$	15 654	16 062	16 954	194	211	254	254	254
	Couverture de l'IC	93,8 %	93,5 %	94,3 %	94,8 %	93,6 %	93,7 %	93,7 %	93,7 %

I'avons mentionné antérieurement, l'imputation par la régression linéaire est généralement biaisée en théorie. Cependant, le biais est faible en raison de la forte dépendance linéaire des données.

3. La méthode (7) ne donne pas de bons résultats pour $t \geq 3$ pour les strates à tirage partiel, parce que l'hypothèse que la propension à répondre dépend de la dernière valeur observée ne tient pas.

4. Les méthodes (8) et (9) donnent de bons résultats, de nouveau en raison de la forte dépendance linéaire de données. Même si ces méthodes utilisent un plus grand nombre de données observées dans l'imputation par la régression, les résultats sont comparables à ceux de la méthode par la régression linéaire proposée.

Tableau 4
Estimations des dépenses totales en R-D (en milliers) d'après les données de la SIRD pour les années 2002 à 2005.
L'erreur-type bootstrap (en milliers) entre parenthèses¹

Méthode	Strates à tirage complet				Strates à tirage partiel			
	t = 2	t = 3	t = 4	t = 2	t = 3	t = 4	t = 2	t = 4
Imputation courante	154 066	156 754	168 015	2 694	2 790	2 782	-	-
Répondants seulement sans correction des pondérations	149 502	148 300	159 822	2 448	2 553	2 419	2 448	2 553
Répondants seulement avec correction des pondérations	166 924	172 419	196 815	2 887	3 219	3 269	2 887	3 219
Censure et imputation par la régression linéaire	154 824	159 394	171 633	2 843	2 997	3 161	2 843	2 997
Imputation par la régression à noyau proposée	154 824	161 414	171 603	2 843	2 997	3 161	2 843	2 997
Imputation par la régression linéaire	154 824	161 414	171 603	2 843	2 997	3 161	2 843	2 997
Imputation par la régression linéaire, valeurs d'après de la dernière valeur	154 824	159 401	172 600	2 843	3 098	3 257	2 843	3 098
Imputées antérieures traitées comme observées	154 824	159 401	172 600	2 843	3 098	3 257	2 843	3 098
Imputation par la régression linéaire fondée sur les données observées courantes et antérieures	154 824	160 205	172 452	2 843	3 168	3 273	2 843	3 168
Imputation par la régression linéaire fondée sur les données observées courantes et antérieures	154 824	160 205	172 452	2 843	3 168	3 273	2 843	3 168

¹ Avertissement : Les valeurs du tableau 4 ne représentent pas nécessairement des estimations nationales parce que nous avons appliqué certaines contraintes aux données afin de respecter notre cadre d'étude.

3.3 Résultats de simulation fondés sur la population de la SIRD

Une étude en simulation supplémentaire a été réalisée en utilisant une population créée à partir des données de la SIRD. La simulation a été exécutée indépendamment pour les classes d'imputation des strates à tirage complet et des strates à tirage partiel. Pour créer la population, nous partons des données de la SIRD en imputant les valeurs manquantes par la méthode courante utilisée pour la SIRD. Soit δ_i , le vecteur des indicateurs de réponse observées ou imputées prise i et y_i le vecteur des valeurs observées ou imputées des dépenses totales en R-D de l'entreprisse i au cours du temps, $i = 1, \dots, n$. Pour la simulation, nous effectuons l'échantillonnage à partir d'une population fondée sur $\{y_i, \delta_i\}$, $i = 1, \dots, n$ comme il suit. Nous commençons par tirer un échantillon de taille n avec remise de y_1, \dots, y_n , puis nous ajoutons un bruit aléatoire normal indépendant de moyenne 0 et d'écart-type 500 à chaque composante de chacun des vecteurs échantillonnés. Toute valeur négative résultante est remplacée par 0. Nous désignons ces dépenses totales en R-D simulées par y_1^*, \dots, y_n^* , où n est défini de la même façon qu'à la section 3.2. Nous désignons les indicateurs de réponse simulés par $\delta_1^*, \dots, \delta_n^*$.

$$P(\delta_n^* = 1 \mid y_1^*, \dots, y_{t-1}^*)$$

$$= \frac{\exp(\beta_{(t)}^0 y_{1,t-1}^* + \beta_{(t)}^1 y_{2,t-1}^* + \dots + \beta_{(t)}^{t-1} y_{t-1,t-1}^*)}{1 + \exp(\beta_{(t)}^0 y_{1,t-1}^* + \beta_{(t)}^1 y_{2,t-1}^* + \dots + \beta_{(t)}^{t-1} y_{t-1,t-1}^*)}$$

Les coefficients $\beta_{(t)}^0, \beta_{(t)}^1, \dots, \beta_{(t)}^{t-1}$ sont fixes tout au long de la simulation et sont obtenus comme étant les coefficients estimés d'après un ajustement initial d'une régression logistique de δ_n sur $(y_{1,t-1}^*, \dots, y_{t-1,t-1}^*)$ pour $t = 1, \dots, n$. Le tableau 5 donne les résultats des simulations pour les estimateurs du total fondés sur 1 000 exécutions et les méthodes (1) à (9) décrites à la section 3.1, où les quantités qui figurent dans le tableau sont définies à la section 3.1. Pour calculer le biais relatif, nous avons obtenu la valeur réelle du total grâce à une exécution préliminaire du modèle de simulation. Plusieurs conclusions tirées de la simulation de la population normale de la section 3.1 s'appliquent dans les conditions décrites ici. Voici un résumé de certaines constatations supplémentaires.

Bien que la régression à noyau soit asymptotiquement valide, dans la présente étude en simulation, le nombre total de sujets est de 2 000 et, selon le tableau 2, les nombres moyens de points de données utilisés dans la régression à noyau sous les schémas $(t, r) = (4, 1)$ et $(4, 2)$ sont de 238 et 152, respectivement, ce qui pourrait ne pas suffire pour la régression à noyau et causer de légers biais dans l'imputation. En revanche, la régression linéaire est plus stable et donne de bons résultats pour une taille d'échantillon telle que 152. Bien qu'en théorie, l'imputation par la régression linéaire donne un biais, celui-ci peut être faible quand $E(y_i | y_{-i}, \dots, y_{i-1})$ est linéaire.

3.2 Application à la SIRD

La SIRD est une enquête annuelle menée auprès d'environ 31 000 entreprises susceptibles de faire de la recherche et du développement. La NSF parraine cette enquête dans le cadre d'un mandat exigeant qu'elle recueille, interprète et analyse des données sur les ressources en sciences et en génie aux États-Unis. L'enquête est menée conjointement par le U.S. Census Bureau et la NSF. Il est demandé aux entreprises sur lesquelles on a l'enquête de fournir des renseignements sur leurs dépenses totales en recherche et développement (R-D) durant l'année civile sur laquelle porte l'enquête. Chaque année, la SIRD est menée de manière déterministe auprès de certaines entreprises en les plaçant dans une strate à tirage complet, puisqu'elles représentent un pourcentage important de l'investissement monétaire total en R-D aux États-Unis. Les autres entreprises qui participent à l'enquête sont échantillonnées chaque année en utilisant un plan d'échantillonnage avec probabilité proportionnelle à la taille (PPT). Des mesures longitudinales sont disponibles pour le noyau d'entreprises qui sont échantillonnées avec certitudes et pour les entreprises de la strate à tirage partiel se trouvant être sélectionnées chaque année. En vue d'illustrer nos méthodes d'imputation, nous nous limitons à n'examiner que les entreprises qui ont été sélectionnées pour l'enquête chaque année de 2002 à 2005 ($T = 4$), et les entreprises qui ont fourni une réponse en 2002. Pour de la documentation sur la SIRD et des tableaux statistiques détaillés, nous renvoyons le lecteur au document intitulé *Research and Development in Industry: 2005*, qui peut être consulté à l'adresse <http://www.nsf.gov/statistics/nsf0319>. D'autres renseignements sur la *Business R&D and Innovation Survey* peuvent être consultés en ligne aux adresses <http://bhs.dev.econ.census.gov/bhs/brdis/> et <http://www.nsf.gov/statistics/srv/industry/about/brdis/>.

Nous divisons les données en deux classes d'imputation. L'une comprend toutes les entreprises contenues dans la strate à tirage complet chacune des quatre années; la seconde comprend toutes les autres entreprises. Dans chaque classe d'imputation, les données sont de la forme (y_i, δ_i) ,

$i = 1, \dots, n_i$, où y_i représente les dépenses totales en R-D de l'entreprise i au temps $t = 1$ (2002), 2 (2003), 3 (2004), 4 (2005). Ici, la taille de l'échantillon est $n = 2\,309$ pour la classe des strates à tirage complet et $n = 1\,039$ pour la classe des strates à tirage partiel. La réponse manquante est non monotone et les pourcentages de réponses manquantes pour 2003, 2004 et 2005 étaient de 10,4 %, 14,0 % et 18,8 % pour la classe des strates à tirage complet et de 15,2 %, 20,7 % et 26,0 % pour la classe des strates à tirage partiel.

Le tableau 4 donne les totaux et les erreurs-types estimés en utilisant les méthodes (2) à (9) décrites à la section 3.1. Comme il est discuté à la fin de la section 2.1, dans chacune des méthodes d'imputation proposées, nous utilisons la régression linéaire quand $r + 1 = t$. Les erreurs-types pré-sentées au tableau 4 ont été calculées par la méthode du bootstrap. Le tableau 4 donne aussi les totaux estimés obtenus quand les données manquantes sont remplacées par les valeurs établies par le Census Bureau afin de produire les tableaux de données publiés officiellement (ces tableaux sont disponibles en ligne à l'adresse http://www.nsf.gov/statistics/pubsubser.cfm?ser_id=26). La méthode utilisée par le Census Bureau pour traiter les données manquantes en vue de produire ces tableaux de données publiés (que nous appelons « méthode courante ») était l'imputation par le ratio pour les entreprises pour lesquelles des données étaient disponibles pour les années précédentes, en utilisant des cellules d'imputation formées par le type d'industrie; nous renvoyons le lecteur à Bond (1994) pour d'autres renseignements. Le tableau 4 donne aussi les dépenses totales en R-D estimées en se basant sur les répondants seulement sans correction de la pondération, qui montrent que ne pas tenir compte des données manquantes introduit un biais dans les estimations. Les méthodes (3) à (9) donnent des résultats comparables, vraisemblablement en raison de la forte dépendance linéaire des données, qui fait en sorte que des méthodes biaisées en théorie présentent un biais négligable. Les totaux estimés fondés sur la méthode courante sont comparables à ceux fondés sur les méthodes proposées pour le cas des strates à tirage complet, mais diffèrent dans le cas des strates à tirage partiel. La méthode combinant la censure et la régression linéaire produit le même écart-type de données écartés par censure n'est pas trop important. Dans la classe d'imputation des strates à tirage complet, 10 % seulement de l'échantillon présente un schéma de non-réponse intermittente et le pourcentage de cas complets est de 72 %. Dans la classe des strates à tirage partiel, 9 % seulement de l'échantillon présente un schéma de non-réponse intermittente et le pourcentage de cas complets est de 66 %.

et 0,5 %, mais il est suffisamment grand pour produire de mauvais résultats de couverture de l'intervalle de confiance apparenté au temps $t = 4$.

5. \bar{Y}_t fondé sur l'imputation par la régression linéaire proposée possède un biais négligeable et une variance plus faible que celle de \bar{Y}_t fondé sur la régression à noyau. La probabilité de couverture de l'intervalle de confiance apparenté est proche du niveau nominal de 95 %.

6. \bar{Y}_t fondé sur l'imputation par la régression à noyau avec indice unidimensionnel est généralement bon, mais un peu moins que l'estimateur fondé sur l'imputation par la régression linéaire.

7. Le biais de \bar{Y}_t fondé sur les méthodes (7) à (9) n'est pas négligeable quand $t = 3$ ou $t = 4$, ce qui entraîne de mauvaises propriétés de l'intervalle de confiance apparenté.

que l'écart-type de la moyenne estimée est également très faible.

2. L'estimateur bootstrap de l'écart-type donne de bons résultats dans tous les cas, même quand l'estimateur de la moyenne est biaisé.

3. Le biais de \bar{Y}_t fondé sur la censure et l'imputation par la régression linéaire est négligeable, de sorte que la probabilité de couverture de l'intervalle de confiance apparenté est proche du niveau nominal de 95 % ; toutefois, son écart-type est grand quand $t = 3$ ou $t = 4$. L'inefficacité de cette méthode est manifestement due au fait d'écartier des données observées provenant de près de 50 % des sujets échantillonnés caractérisés par une non-réponse intermittente. Sa performance empirique à mesure que t augmente.

4. Le biais relatif de \bar{Y}_t fondé sur l'imputation par la régression à noyau proposée est compris entre 0,0 %

Tableau 3 Résultats de simulation pour l'estimation de la moyenne (population normale)

Méthode	Quantité	$t = 2$	$t = 3$	$t = 4$
Données complètes	$E-T$ $E-T_{boot}$ Biais relatif	0 % 0,0221 0,0223	0 % 0,0223 0,0223	0 % 0,0221 0,0224
Répondants seulement	$E-T$ $E-T_{boot}$ Biais relatif	12,8 % 0,0282 6,8 %	94,9 % 0,0276 94,4 %	3,5 % 0,0272 95,4 %
Imputation par la régression linéaire	Couverture de l'IC	95,1 %	92,5 %	88,6 %
	$E-T$ $E-T_{boot}$ Biais relatif	0,0 % 0,0275 0,0 %	0,0 % 0,0286 0,1 %	0,0 % 0,0279 0,0 %
	proposée	0,0275	0,0287	0,0293
Imputation par la régression à noyau avec indice unidimensionnel	Couverture de l'IC	95,1 %	93,8 %	95,7 %
	$E-T$ $E-T_{boot}$ Biais relatif	0,0 % 0,0275 0,0 %	0,0 % 0,0288 0,4 %	0,4 % 0,0279 0,0 %
	proposée	0,0276	0,0288	0,0288
Imputation par la régression à noyau avec dépendance à l'égard de la dernière valeur	Couverture de l'IC	95,1 %	92,5 %	88,6 %
	$E-T$ $E-T_{boot}$ Biais relatif	0,0 % 0,0288 0,0 %	0,0 % 0,0295 84,2 %	0,0 % 0,0263 86,2 %
	proposée	0,0276	0,0260	0,0246
Imputation par la régression linéaire avec valeurs imputées antérieurement traitées comme observées	Couverture de l'IC	95,1 %	59,7 %	76,0 %
	$E-T$ $E-T_{boot}$ Biais relatif	0,0 % 0,0275 0,0 %	1,6 % 0,0261 1,6 %	0,8 % 0,0241 0,8 %
	proposée	0,0276	0,0260	0,0246
Imputation sur les données observées courantes et antérieures	Couverture de l'IC	95,1 %	59,0 %	76,1 %
	$E-T$ $E-T_{boot}$ Biais relatif	0,0 % 0,0275 0,0 %	1,6 % 0,0261 1,6 %	0,8 % 0,0242 0,8 %
	proposée	0,0276	0,0261	0,0246

3.1 Résultats de simulation pour une population normale

Nous avons exécuté une étude par simulation en utilisant une population normalement distribuée Y_1, \dots, Y_n , $n = 2\,000$ et $T = 4$. Nous avons en outre utilisé une seule classe d'imputation et un plan d'échantillonnage aléatoire simple avec remise. Dans la simulation, les y_t ont été générés de manière indépendante à partir de la loi normale multivariée de vecteur de moyennes (1,33 ; 1,94 ; 2,73 ; 3,67) et de matrice de covariance ayant une structure AR(1) avec coefficient de corrélation de 0,7 et variance unitaire ; au temps $t = 1$, toutes les données ont été observées ; les données manquantes au temps $t = 2, 3, 4$ ont été générées conformément à

$$P(\delta_t = 1 | Y_1, \dots, Y_{t-1}, \delta_{t-1}) = 1 - \Phi\left(0,6\right) - \sum_{j=1}^{t-1} Y_j Y_{\delta_{t-1}}^{j-1}$$

où Φ est la fonction de répartition normale centrée réduite. Les probabilités non conditionnelles des schémas de non-réponse sont données au tableau 2.

Tableau 2
Probabilités des schémas de non-réponse dans l'étude en simulation (population normale)

Schéma		Probabilité	
Monotone	(1,0,0,0)	0,062	total = 0,181
	(1,1,0,0)	0,043	
	(1,1,1,0)	0,076	
Intermittente	(1,0,0,1)	0,113	total = 0,494
	(1,0,1,0)	0,071	
	(1,0,1,1)	0,186	
	(1,1,0,1)	0,124	
Complète	(1,1,1,1)	0,325	

Pour les comparaisons, nous avons inclus neuf estimateurs de la moyenne de y_t , à savoir les moyennes d'échantillons fondées sur 1) les données complètes (utilisée comme norme de référence), 2) les répondants avec poids corrigés en supposant que la probabilité de réponse est la même dans chaque classe d'imputation, 3) la censure et l'imputation par la régression linéaire, qui consiste à écarter d'abord toutes les observations d'un sujet après la première valeur manquante pour créer un ensemble de données à « non-réponse monotone », puis à appliquer l'imputation par la régression linéaire comme il est décrit dans Paik (1997), 4) l'imputation par la régression linéaire proposée, 5) l'imputation par la régression à noyau avec indice unidimensionnel proposée, en utilisant la régression inverse par tranches pour obtenir \hat{y}_t ; 7) l'imputation par la régression à noyau proposée dans

Xu et coll. (2008) fondée sur la propension à répondre dépendant de la dernière valeur, 8) l'imputation par la régression linéaire fondée sur une régression des réponses au temps t sur les valeurs observées et imputées aux points dans le temps 1, ..., $t - 1$ (en traitant les valeurs imputées comme des valeurs observées), 9) l'imputation par la régression linéaire fondée sur une régression des réponses au temps t sur les données observées pour les unités ayant le même schéma de données manquantes aux points dans le temps 1, ..., $t - 1$.

La méthode (2) ne tient simplement pas compte des non-répondants et est donc biaisée et inefficace. Sous l'hypothèse de propension à répondre (1), les méthodes (7) à (9) sont également biaisées pour $t \geq 3$, parce que la méthode (7) requiert l'hypothèse de dépendance à l'égard de la dernière valeur qui est plus forte que l'hypothèse (1), la méthode (8) traite les valeurs imputées antérieurement comme des valeurs observées dans la régression, et la méthode (9) requiert la condition qui suit, qui n'est pas vérifiée sous (1) :

$$E(Y_t | Y_1, \dots, Y_{t-1}, \delta_{t-1} = j_1, \dots, \delta_{t-1} = j_{t-1}, \delta_t = 0) = E(Y_t | Y_1, \dots, Y_{t-1}, \delta_{t-1} = j_1, \dots, \delta_{t-1} = j_{t-1}, \delta_t = 1)$$

où (j_1, \dots, j_{t-1}) est un schéma de données manquantes fixe. Enfin, comme il est discuté à la section 2.3, la méthode (5) est également biaisée pour $t \geq 3$, puisque la régression linéaire n'est pas un modèle exactement correct. Cependant, les méthodes (5), (8) et (9) peuvent quand même donner de bons résultats quand les biais ne sont pas importants, parce que l'emploi d'un modèle plus simple et de plus de données dans la régression pour l'imputation peut compenser la perte due à l'imputation biaisée. En outre, toute hypothèse concernant la propension à répondre peut n'être vérifiée qu'approximativement et il est souhaitable d'étudier empiriquement diverses méthodes dans toute application particulière. Pour le cas où $r = t - 1$, nous appliquons l'imputation par la régression linéaire comme il est exposé à la section 2.1. Donc, les méthodes (3) à (6), (8) et (9) donnent toutes de bons résultats quand $t = 2$.

Le tableau 3 donne (sur la base de 1 000 exécutions de la simulation) le biais relatif et l'écart-type $(E - T)$, de l'estimateur de la moyenne, la moyenne de $E - T$, l'estimateur bootstrap de l'écart-type fondé sur 200 répétitions bootstrap, et la probabilité de couverture de l'intervalle de confiance (IC) à 95 % approximatif obtenu en utilisant l'estimateur ponctuel $\pm 1,96 \times E - T$. Les résultats du tableau 3 se résument comme il suit.

sondage. À tout temps t , soit $y_t^n = y^n$ quand $\delta^n = 1$ et y_t^n la valeur imputée en utilisant l'une des méthodes de la section 2 quand $\delta^n = 0$. Le total de population finie et la moyenne de y_t peuvent être estimés par

$$\bar{y}_t = \sum_{i \in S} w_i y_t^i \quad \text{et} \quad \bar{Y}_t = \sum_{i \in S} w_i \bar{y}_t^i / \sum_{i \in S} w_i \quad (9)$$

respectivement, où w_i est le poids de sondage construit de façon que, s'il n'y a pas de non-réponse, \bar{Y}_t soit un estimateur sans biais du total de population finie au temps t sous le plan d'échantillonnage probabiliste. La moyenne de superpopulation de y_t peut alors être estimée par \bar{Y}_t . Notons que $\sum_{i \in S} w_i$ est un estimateur sans biais de la taille de la population finie N et, pour certains plans d'échantillonnage simples, il est exactement égal à N .

Les poids de sondage devraient également être utilisés dans l'ajustement de la régression pour l'imputation. Sous les mêmes conditions que celles données dans Cheng (1994), \bar{Y}_t ou \bar{Y}_t^* fondé sur l'imputation par la régression à noyau ou par la régression à indice unidimensionnel est convergent et asymptotiquement normal à mesure que la taille de l'échantillon tend vers ∞ . Les conditions requises et les preuves peuvent être consultées dans Xu (2007).

Si nous appliquons une méthode d'imputation par la régression linéaire telle qu'elle est exposée à la section 2.3, la moyenne estimée résultante au temps t peut être asymptotiquement biaisée. Ce biais est faible s'il est possible de bien approximer la fonction w_t^* par une fonction linéaire dans l'étendue des valeurs des données. Par ailleurs, l'imputation par la régression à noyau ou par la régression à noyau avec indice unidimensionnel peut nécessiter un échantillon de beaucoup plus grande taille que l'imputation par la régression linéaire. Par conséquent, les propriétés globales de la moyenne estimée en utilisant l'imputation par la régression linéaire peuvent demeurer meilleures, comme l'indiquent les résultats de simulation présentés à la section 3.

2.5 Estimation de la variance

Afin d'évaluer l'exactitude statistique ou l'inférence, comme la construction d'un intervalle de confiance pour la moyenne de y_t au temps t , nous avons besoin des estimateurs de variance de \bar{Y}_t ou \bar{Y}_t^* fondés sur des données imputées. Étant donné la complexité de la procédure d'imputation, il est difficile d'obtenir des formules explicites pour la variance de \bar{Y}_t ou \bar{Y}_t^* . Nous considérons alors la méthode du bootstrap (Efron 1979). Un bootstrap correct peut être obtenu en répétant le processus d'imputation dans chacun des échantillons bootstrap (Shao et Sitter 1996). Soit $\hat{\theta}$ l'estimateur examiné. Une procédure bootstrap peut être exécutée comme il suit.

régression.

1. Tirer de S un échantillon bootstrap sous forme d'échantillon aléatoire simple de même taille avec remise parmi l'ensemble de sujets échantillonnés.
 2. Utiliser les poids de sondage, les indicateurs de réponse et les données observées provenant de l'ensemble de données original pour les unités de l'échantillon bootstrap pour former un ensemble de données bootstrap. Appliquer la procédure d'imputation proposée aux données bootstrap. Calculer l'analogue bootstrap $\hat{\theta}^*$ de $\hat{\theta}$.
 3. Indépendamment, répéter B fois les étapes qui précèdent pour obtenir $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$. La variance d'échantillon de $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ est l'estimateur de variance bootstrap pour $\hat{\theta}$.
- Dans l'application, chaque $\hat{\theta}^{*b}$ peut être calculé en utilisant les b^e données bootstrap $(y^{(i)b}, w_i^{(b)}), i \in S$, où $w_i^{(b)} = w_i$ est multiplié par le nombre de fois que l'unité i figure dans le b^e échantillon bootstrap. Notons que le même $w_i^{(b)}$ peut être utilisé pour toute les variables d'intérêt, plutôt que pour y_t seulement.
- ### 3. Résultats empiriques
- Nous étudions \bar{Y}_t ou \bar{Y}_t^* dans (9) obtenu à chaque point t dans le temps en se fondant sur les méthodes d'imputation proposées. Nous considérons d'abord une simulation comparative. Nous présentons une application aux données de la SIRD. Pour examiner les propriétés des méthodes proposées pour la population créée en utilisant les données de la SIRD, nous terminons par une simulation portant sur une régression linéaire non paramétrique requises, nous utilisons la fonction *loess* de R avec les valeurs par défaut des paramètres, qui ajuste une surface polynomiale locale pour une ou plusieurs variables explicatives. Les régressions linéaires requises sont ajustées facilement en R en utilisant la fonction *lm*. Nos implémentations des méthodes proposées comprennent la vérification de l'erreur (par exemple veiller à ce que le nombre de points soit suffisant pour l'ajustement de la régression à chaque étape), ce qui est particulièrement important dans des conditions de bootstrap et de simulation où les méthodes d'imputation sont répétées de nombreuses fois et où chaque itération ne peut pas être examinée manuellement. Nous avons choisi comme option par défaut l'imputation d'une moyenne globale dans le cas où le nombre de points n'était pas suffisant pour ajuster une

explicatives dans la régression, car l'équation (6) n'est pas vérifiée si certaines des valeurs y_1, \dots, y_s sont imputées. Alors que l'on utilise toutes les données observées à $E(y_t)$, on se sert de certaines données observées à un temps $t < t_s$ mais pas de toutes pour l'imputation, afin d'éviter des biais sans non-réponse non ignorable. La situation est différente en cas de non-réponse observées, où habituellement toutes les données observées antérieures peuvent être utilisées dans l'imputation par la régression.

2.3 Régression pour l'imputation

Dans (5), les espérances conditionnelles dépendent non seulement de la distribution de y , mais aussi de la proportion à répondre. Même si $E(y_t | y_1, \dots, y_{t-1})$ est linéaire, les espérances conditionnelles figurant dans (5) ne le sont pas forcément, ce qui diffère du cas où $r + 1 = t$ considéré à la section 2.1. Le résultat (10) à l'annexe en est un exemple.

Si nous ne disposons pas d'un modèle paramétrique approprié pour $\phi_{t,r}$, nous pouvons appliquer la régression par la méthode du noyau, ou régression à noyau, non paramétrique donnée dans Cheng (1994) pour obtenir $\hat{\phi}_{t,r}$. Puisque la variable dépendante $(y_1^{(r)}, \dots, y_p^{(r)})$ est multivariée quand $r \geq 2$, la régression à noyau présente toutefois une grande variabilité, à moins que le nombre de sujets échantonnés dans la catégorie définie par $\delta_n = \dots = \delta_{(r+1)} = 1$ soit très grand. Ce problème porte le nom de malediction de la dimension.

Donc, nous considérons les options qui suivent sous l'hypothèse supplémentaire que la dépendance de δ_t à l'égard de y_1, \dots, y_{t-1} a lieu par la voie d'une combinaison linéaire de y_1, \dots, y_{t-1} . C'est-à-dire

$$P(\delta_t = 1 | y_1, \dots, y_{t-1}, \delta_{t-1}, \dots, \delta_1) = \Psi \left(\sum_{l=1}^L \delta_{t-1}^{l_1} \dots \delta_{t-1}^{l_L} y_{t-1}^{l_L} \right), \quad (7)$$

où $\delta_{t-1}^{l_1} \dots \delta_{t-1}^{l_L}$, $l = 1, \dots, t-1$, sont les paramètres inconnus qui dépendent de $\delta_1, \dots, \delta_{t-1}$ et Ψ est une fonction inconnue dans l'intervalle $[0, 1]$. Sous (7), nous montrons à l'annexe que

$$E(y_t | z, \delta_t = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) \\ = E(y_t | z, \delta_t = \dots = \delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) \\ = 1, \dots, t-2, t-1, \dots, T, \quad (8)$$

où $z_r = \sum_{l=1}^L y_{t-1}^{l_1} y_{t-1}^{l_2} \dots y_{t-1}^{l_L}$ et $y_{t-1}^{l_1} \dots y_{t-1}^{l_L} = \delta_r = 1$. Donc, pour imputer les valeurs des non-répondants, nous pouvons conditionner sur la combinaison linéaire z_r et utiliser (8) au lieu de conditionner sur y_1, \dots, y_t et d'utiliser (5).

Soit $\psi_{t,r}(z_r)$ la fonction définie à la deuxième ligne de (8). Souhaitons que $\psi_{t,r}$ n'est pas nécessairement identique à $\phi_{t,r}$. S'il existe une forte relation linéaire entre y_t et y_1, \dots, y_p , il se pourrait que $\psi_{t,r}$ soit approximativement linéaire de sorte que nous pouvons ajuster une régression linéaire pour obtenir un estimateur $\hat{\psi}_{t,r}$. En théorie, cette méthode contient un biais quand $\psi_{t,r}$ n'est pas linéaire. Si $\gamma_r = (y_{r,1}, \dots, y_{r,p})$ est connu, nous pouvons appliquer une régression à noyau unidimensionnelle pour obtenir un estimateur $\hat{\psi}_{t,r}$, en utilisant l'indice unidimensionnel z_r . Puisque γ_r est inconnu, nous devons d'abord l'estimer par $\hat{\gamma}_r$, puis obtenir $\hat{\psi}_{t,r}$ en exécutant la régression à noyau unidimensionnelle avec remplacement de γ_r par $\hat{\gamma}_r$. Par exemple, on peut recourir à la régression inverse par tranches (Duan et Li 1991) pour obtenir $\hat{\gamma}_r$. Cependant, ce genre de méthode non paramétrique est parfois inefficace. S'il existe une forte relation linéaire entre y_t et y_1, \dots, y_p , nous pouvons appliquer la régression linéaire pour obtenir $\hat{\gamma}_r$. Quoiqu'il en soit, nous utilisons $y_{r,1}, \dots, y_{r,p}$ avec $\delta_n = \dots = \delta_{(r+1)} = 1$ comme variables explicatives et les valeurs $y_n^{(r)}$ imputées comme réponses dans tout type d'ajustement de la régression. Après avoir obtenu $\hat{\psi}_{t,r}$ et $\hat{\gamma}_r = (\hat{\gamma}_{r,1}, \dots, \hat{\gamma}_{r,p})$, nous imputons une valeur $y_n^{(r)}$ manquante par $\hat{y}_n^{(r)} = \hat{\psi}_{t,r}(\hat{\gamma}_{r,1}, y_{r,1} + \dots + \hat{\gamma}_{r,p}, y_{r,p})$.

Ces méthodes peuvent aussi être appliquées au cas où $r = t-1$ si $E(y_t | y_1, \dots, y_{t-1})$ n'est pas linéaire. En théorie, des estimateurs tels que les moyennes estimées en se fondant sur l'imputation par la régression à noyau ou par la régression à noyau avec indice unidimensionnel sont asymptotiquement sans biais, mais ils ne sont peut-être pas meilleurs que ceux fondés sur l'imputation par la régression linéaire quand le nombre de sujets échantonnés dans chaque catégorie (t, r) n'est pas très grand. Les propriétés des moyennes estimées en utilisant l'imputation par la régression linéaire, par la régression à noyau et par la régression à noyau avec indice unidimensionnel sont examinées par simulation à la section 3.

2.4 Estimation

Nous considérons l'estimation du total de population finie ou de la moyenne de y_t à chaque point t fixé dans le temps, ce qui est souvent l'objectif principal d'une étude par

2.2 Imputation pour des sujets dont la première

valeur manquante a lieu au temps $r + 1 < t$

L'imputation pour un sujet dont la première valeur manquante se produit au temps $r + 1 < t$ est plus compliquée et diffère de celle applicable au cas de non-réponse monotone. En effet, quand $r + 1 < t$ et que la non-réponse est monotone,

$$E(y_i | y_{i1}, \dots, y_{ir}, \delta_i = \dots = \delta_r = 1, \delta_i = 0) \\ = E(y_i | y_{i1}, \dots, y_{ir}, \delta_i = \dots = \delta_r = 1, \delta_i = 1) \quad (4)$$

$$r = 1, \dots, t - 2, \quad t = 2, \dots, T. \quad (4)$$

tandis que (4) n'est pas vérifiée quand la non-réponse est non monotone (voir la preuve à l'annexe). D'où, nous devons spécifier des modèles différents pour les sujets dont la première valeur manquante se produit à $r + 1 < t$. Nous montrons à l'annexe que, quand $r + 1 < t$,

$$E(y_i | y_{i1}, \dots, y_{ir}, \delta_i = \dots = \delta_r = 0, \delta_i = 0) \\ = E(y_i | y_{i1}, \dots, y_{ir}, \delta_i = \dots = \delta_r = 1, \delta_i = 0) \quad (5)$$

$$r = 1, \dots, t - 2, \quad t = 2, \dots, T. \quad (5)$$

Expliquons maintenant comment utiliser (5) pour imputer les valeurs manquantes à un point dans le temps fixe t . Soit $\phi_{i1}, \dots, \phi_{ir}$ la quantité à la première ligne de (5). Si $\phi_{i1}, \dots, \phi_{ir}$ est connue, y_i peut être imputée par $\phi_{i1}, \dots, \phi_{ir}$. Sinon, elle doit être estimée en se fondant sur (5). Contrairement au modèle (2) ou (4), l'espérance conditionnelle à la deuxième ligne de (5) est conditionnelle à une valeur manquante y_{ir} ($\delta_i = 0$), même si les valeurs y_{i1}, \dots, y_{ir} sont observées. Si nous exécutons l'imputation séquentiellement, conformément à $r = t - 1, t - 2, \dots, 1$, alors, pour un temps donné $r < t - 1$, les valeurs y_i manquantes pour les sujets dont la première valeur manquante a eu lieu au point dans le temps $r + 2$ sont déjà imputées par la méthode décrite à la présente section ou à la section 2.1. Nous pouvons ajuster une régression de la valeur imputée y_{ir} sur les valeurs observées y_{i1}, \dots, y_r en utilisant les données provenant de tous les sujets pour lesquels on dispose de la valeur y_i déjà imputée (utilisées comme réponses), des valeurs y_{i1}, \dots, y_r observées (utilisées comme variables explicatives) et de $\delta_{i,r+1} = 1$. Une fois qu'un estimateur $\hat{\phi}_{ir}$ est obtenu, une valeur manquante y_{ir}^u avec une première valeur manquante au temps $r + 1$ est alors imputée par $\hat{y}_{ir}^u = \hat{\phi}_{ir}(y_{i1}, \dots, y_{ir})$. Considérons de nouveau le cas où $t = 3$ ou 4 et le tableau 1. Après la première étape pour $t = 3$ décrite à la section 2.1, à la deuxième étape, nous imputons les valeurs manquantes avec $r = 1$ selon un schéma (1,0,0). Conformément au modèle d'imputation (5), nous ajustons une régression en utilisant des données présentant un schéma (1,1,0) indiquées par + (utilisées comme variables explicatives) et \otimes (valeurs imputées antérieurement utilisées) et (1,1,0) indiquées par - (valeurs imputées antérieurement utilisées). Alors, les valeurs imputées (indiquées par \otimes) sont obtenues d'après la régression ajustée en utilisant les données indiquées par * comme variables explicatives. À l'étape 3, les données indiquées par * comme variables explicatives. Pour $t = 4$, après la première étape décrite à la section 2.1, à la deuxième étape ($r = 2$), nous ajustons une régression en utilisant des données présentant un schéma (1,1,1,0) en utilisant des données (indiquées par + (utilisées comme variables explicatives) et \otimes (valeurs imputées antérieurement utilisées) et \otimes (valeurs imputées antérieurement utilisées) comme réponses). Alors, les valeurs imputées (indiquées par \otimes) au temps $t = 4$ présentant un schéma (1,1,1,0) sont obtenues d'après la régression ajustée en utilisant les données indiquées par * comme variables explicatives. À l'étape 3, les données indiquées par * comme variables explicatives. Bien qu'au temps t l'imputation doit être exécutée séquentiellement suivant $r = t - 1, \dots, 1$, l'imputation pour différents points dans le temps peut être effectuée dans n'importe quel ordre. On peut le constater en examinant l'exemple du tableau 1, où les valeurs imputées à $t = 3$ n'interviennent pas dans le processus d'imputation à $t = 4$ ou inversement, quoique certaines données observées seront utilisées à plusieurs reprises dans l'ajustement de la régression. Lorsque les données sont fournies en fonction du temps, il est naturel d'imputer les non-répondants dans l'ordre $t = 2, \dots, T$.

Pourquoi pouvons-nous utiliser des valeurs imputées antérieurement comme réponses dans l'estimation de la fonction de régression $\phi_{i,r}$ quand $r < t - 1$? Pour t et $r < t - 1$ donné, une valeur imputée antérieurement avec la première valeur manquante à $s + 1 > r + 1$ est un estimateur de

$$y_i = E(y_i | y_{i1}, \dots, y_{is}, \delta_i = \dots = \delta_s = 1, \delta_i = 0) \\ = E(y_i | y_{i1}, \dots, y_{is}, \delta_i = \dots = \delta_s = 1, \delta_i = 0).$$

En vertu de la propriété d'espérance conditionnelle et de (5),

$$E[E(y_i | y_{i1}, \dots, y_{is}, \delta_i = \dots = \delta_{s+1} = 1, \delta_i = 0) | \\ y_{i1}, \dots, y_{ir}, \delta_i = \dots = \delta_{r+1} = 1, \delta_i = 0] \\ = E(y_i | y_{i1}, \dots, y_{ir}, \delta_i = \dots = \delta_{r+1} = 1, \delta_i = 0) \\ = E(y_i | y_{i1}, \dots, y_{ir}, \delta_i = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_i = 0). \quad (6)$$

D'où y_i et y_i ont la même espérance conditionnelle, sachant $y_{i1}, \dots, y_{ir}, \delta_i = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_i = 0$. Par conséquent, l'utilisation des valeurs imputées antérieurement comme réponses dans la régression produit un estimateur valide de $\phi_{i,r}$. Notons que les valeurs imputées antérieurement ne devraient pas être utilisées comme variables

ajuster la régression.

Quel type de régression pouvons-nous ajuster pour obtenir \hat{y}^n ? Nous montrons à l'annexe que, si (1) est vérifiée et que $E(y_i | y_1, \dots, y_{i-1})$ est linéaire en y_1, \dots, y_{i-1} pour tout i en l'absence de non-réponse, alors

$$E(y_i | y_1, \dots, y_{i-1}, \delta_1 = \dots = \delta_{i-1} = 1)$$

est linéaire en y_1, \dots, y_{i-1} (3)

(3) $\gamma_1, \dots, \gamma_{l-1}$ est linéaire en

$$(1 = \delta_{i-1} = \dots = \delta_1 \delta_{i-1}, \chi_{i-1}, \dots, \chi_1 | \chi'_1)(E$$

Quel type de régression pouvons-nous ajuster pour obtenir $\gamma_j^{(j)}$? Nous montrons à l'annexe que, si (1) est vérifiée et que $E(\gamma_j^{(j)} | \gamma_1^{(j)}, \dots, \gamma_{j-1}^{(j)})$ est linéaire en $\gamma_1^{(j)}, \dots, \gamma_{j-1}^{(j)}$ pour tout j en l'absence de non-réponse, alors

c'est-à-dire le schéma (1,1,0). Conformément au modèle d'imputation (2), nous ajustons une régression en utilisant des données ayant un schéma (1,1,1) indiquées par + (utilisées comme variables explicatives) et \times (utilisées comme réponses). Puis, les valeurs imputées (indiquées par \circ) sont obtenues à partir de la régression ajustée en utilisant les données indiquées par * comme variables explicatives. Pour $t = 4$ et $r = 3$, une imputation selon le schéma (1,1,1,0) peut être effectuée de manière similaire en utilisant des données présentant un schéma (1,1,1,1,1,0)

Illustration du processus d'imputation

Schéma												Schéma							
Étape 1 : $r = 3, t = 4$						Étape 2 : $r = 2, t = 4$						Étape 3 : $r = 1, t = 4$							
Temps			Temps			Temps			Temps			Temps			Temps				
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
*	*	*	○	+	+	*	⊗	*	+	+	*	*	+	+	*	○	⊗	⊗	○
+	+	×				⊗													
(1,0,1)	(1,1,1)	(1,1,0)	(1,0,0)	(1,0,1)	(1,1,0)	(1,0,0)	(1,1,1)	(1,0,1)	(1,1,0)	(1,0,0)	(1,1,1)	(1,0,1)	(1,1,0)	(1,0,0)	(1,1,1)	(1,0,1)	(1,1,0)	(1,0,0)	(1,1,1)

* : données observées utilisées dans l'ajustement de la régression comme réponses.

Statistique Canada, N° 12-001-X au catalogue

non-réponse à la période de référence $t = 1$. La propension à répondre dépend des valeurs antérieures si

$$(I) \quad T_{i, \dots, j} = t_{i, \dots, j} g^{i-1} g^{i-1} \dots g^{i-1} \chi \mid I = j g) D =$$

ou F est calculée par rapport à la superpopulation. Quand la non-réponse est monotone, la propension à répondre dépendante des valeurs antérieures devient ignorable (Little et Rubin 2002), puisque nous observons toutes les valeurs antérieures ou savons avec certitude que y_i manque si la valeur manquant à la période $t - 1$, et nous pouvons utiliser une méthode d'imputation par la régression linéaire proposée par Paik (1997). Par contre, quand la non-réponse n'est pas monotone, la propension à répondre dépendante des valeurs antérieures n'est pas ignorable, parce que l'indicateur de réponse au temps t dépend statistiquement des valeurs antérieures de la variable étudiée, dont certaines pourraient ne pas être observées. Dans ce cas, la méthode de Paik ne s'applique pas.

2.1 Imputation pour des sujets dont la première valeur manquante a lieu au temps t

Soit $t < 1$ un point fixe dans le temps et $r + 1$ le point dans le temps auquel la première valeur manquante de y se produit. Quand $r + 1 = t$, c'est-à-dire un sujet dont la première valeur manquante se produit au temps t , la procédure d'imputation que nous proposons est la même que celle pour le cas de la non-réponse monotone (Paik 1997). Cependant, nous devons fournir une justification, puisque nous avons une propension à répondre différente. Nous montrons à l'annexe que, sous l'hypothèse (1),

$$(2) \quad \begin{aligned} & 1 = \delta^{l_1} \delta^{l_2} \cdots \delta^{l_{l_1+l_2+\dots+l_{l_1+l_2+\dots+l_{l_1+l_2+\dots}}}} \\ & 0 = \delta^{l_1} \delta^{l_2} \cdots \delta^{l_{l_1+l_2+\dots+l_{l_1+l_2+\dots+l_{l_1+l_2+\dots}}}} \end{aligned}$$

où E est l'espérance par rapport à la superpopulation. Désignons la quantité à la première ligne de (2) par $\phi_{i-1}^{(j-1)}(Y_j^{i-1})$, qui est l'espérance conditionnelle d'une valeur de Y_j^{i-1} , manquant sachant les valeurs observées Y_j^1, \dots, Y_j^{i-1} . Si $\phi_{i-1}^{(j-1)}$ est connue, une valeur imputée naturelle pour Y_j^i est $\phi_{i-1}^{(j-1)}(Y_j^1, \dots, Y_j^{i-1})$. Cependant, $\phi_{i-1}^{(j-1)}$ est habituellement inconnue. Puisque $\phi_{i-1}^{(j-1)}$ ne peut pas être estimée par la régression de Y_j^i sur Y_j^1, \dots, Y_j^{i-1} en se fondant sur des données provenant de sujets non lesquels des valeurs de Y_j^i manquent, nous devons utiliser (2), c'est-à-dire le fait que $\phi_{i-1}^{(j-1)}$ est égale à la quantité figurant à la deuxième ligne de (2), qui est l'espérance conditionnelle d'une valeur Y_j^i observée sachant les valeurs observées Y_j^1, \dots, Y_j^{i-1} et qui utilise l'estimée par la régression de Y_j^i sur Y_j^1, \dots, Y_j^{i-1} en utilisant les données provenant de tous les sujets non lesquels on dispose de la valeur observée Y_j^i et des valeurs observées Y_j^1, \dots, Y_j^{i-1} . Notons que (2) est l'équivalent de (5) dans Xu et coll. (2008) sous l'hypothèse de dépendance à

2. Méthodologie

Le but du présent article est de proposer une méthode d'imputation pour les données longitudinales en présence de non-réponse non monotone sous hypothèse de propension à répondre dépendante des valeurs antérieures décrite par Little (1995) : au point t dans le temps, la propension à répondre dépend des valeurs de la variable étudiée aux points dans le temps antérieurs à t . Cette hypothèse concernant la propension à répondre est plus faible que celle formulée dans Xu et coll. (2008) et diffère de celles figurant dans Vansteelandt et coll. (2007). Nous considérons une imputation qui ne requiert pas la spécification d'un modèle de la propension à répondre. L'imputation est utilisée fréquemment pour remplacer les valeurs manquantes dans les problèmes de sondage (Kallott et Kasprzyk 1986). Une fois que toutes les valeurs manquantes sont imputées, les estimations des paramètres sont calculées en se servant des moyennes estimées pour les données complètes en traitant les valeurs imputées comme des observations. La méthodologie d'imputation et d'estimation proposée, y compris une méthode de bootstrap pour l'estimation de la variance, est présentée à la section 2. Afin d'examiner les propriétés en échantillon fini de la méthode proposée, nous présentons certains résultats de simulation à la section 3. Nous décrivons aussi une application de la méthode proposée à la SIRD. Enfin, à la dernière section, nous présentons certaines conclusions.

(Considérons l'approche assistée par modèle pour des données d'enquête échantillonnelles à partir d'une population finie F . Nous supposons que la population F est divisée en un nombre fixe de classes d'imputation, qui sont habituellement des unions de certaines strates. Dans chaque classe d'imputation, la variable étudiée d'une unité de la population provient d'une superpopulation. Soit y_t la variable étudiée au point dans le temps t , $t = 1, \dots, T$, $y = (y_1, \dots, y_T)$, l'indicateur précisant si y_t est observé, et $\delta = (\delta_1, \dots, \delta_T)$. Puisque l'imputation est effectuée indépendamment dans chaque classe d'imputation, pour simplifier la notation nous supposons à la présente section qu'il n'existe qu'une seule classe d'imputation.

Dans tout l'exposé, nous considérons que la non-réponse est non monotone et nous supposons qu'il n'y a pas de

Imputation pour la non-réponse non monotone dans le *Survey of Industrial Research and Development*

Jun Shao, Martin Klein et Jing Xu¹

Résumé

Dans les études longitudinales, la non-réponse est souvent de nature non monotone. Dans le cas de la *Survey of Industrial Research and Development* (SIRD), il est raisonnable de supposer que le mécanisme de non-réponse dépend des valeurs antérieures, en ce sens que la propension à répondre au sujet d'une variable étudiée au point t dans le temps dépend de la situation de réponse ainsi que des valeurs observées ou manquantes de la même variable aux points dans le temps antérieurs. Puisque cette non-réponse n'est pas ignorable, l'approche axée sur la vraisemblance paramétrique est sensible à la spécification des modèles paramétriques s'appuyant sur la distribution conjointe des variables à différents points dans le temps et sur le mécanisme de non-réponse. La non-réponse non monotone limite aussi l'application des méthodes de pondération par l'inverse de la propension à répondre. En écartant toutes les valeurs observées auprès d'un sujet après la première valeur manquante pour ce dernier, on peut créer un ensemble de données présentant une non-réponse monotone ignorable, puis appliquer les méthodes établies pour la non-réponse ignorable. Cependant, l'abandon de données observées ne pas souhaitable et peut donner lieu à des estimateurs inefficaces si le nombre de données écartées est élevé. Nous proposons d'imputer les réponses manquantes par la régression au moyen de modèles d'imputation créés prudemment sous le mécanisme de non-réponse dépendante des valeurs antérieures. Cette méthode ne requiert l'ajustement d'aucun modèle paramétrique sur la distribution conjointe des variables à différents points dans le temps ni sur le mécanisme de non-réponse. Les propriétés des moyennes estimées en appliquant la méthode d'imputation proposée sont examinées en s'appuyant sur des études en simulation et une analyse empirique des données de la SIRD.

Mots clés : Bootstrap ; modèle d'imputation ; régression à noyau ; ne manquant pas au hasard ; étude longitudinale ; dépendante des valeurs antérieures.

1. Introduction

Les études longitudinales, dans lesquelles des données sont recueillies auprès de chaque sujet échantillonné à plusieurs points dans le temps, sont très courantes dans des domaines de recherche tels que la médecine, la santé des populations, l'économie, les sciences sociales et les enquêtes par sondage. Habituellement, l'analyse statistique des données d'une enquête vise à estimer la moyenne d'une variable étudiée, ou à faire une inférence sur cette moyenne, à chaque point dans le temps. La non-réponse ou les données manquantes pour la variable étudiée représentent un obstacle sérieux à l'exécution d'une analyse statistique valide, parce que la propension à répondre de la variable étudiée. La non-réponse est monotone si, quand une valeur manque à un point t dans le temps, toutes les futures valeurs au temps $s > t$ manquent. Nous nous concentrons sur la non-réponse non monotone, qui est fréquente dans les enquêtes longitudinales. Dans la *Survey of Industrial Research and Development* (SIRD), menée conjointement par le U.S. Census Bureau et la U.S. National Science Foundation (NSF), par exemple, une entreprise pourrait ne pas indiquer ses dépenses de recherche et de développement à l'année $t - 1$, mais le faire à l'année t . Pour simplifier,

nous faisons référence à la SIRD au temps présent tout au long de l'exposé, mais nous tenons à signaler qu'à partir de 2008, elle a été remplacée par la *Business R&D and Innovation Survey*.

Certaines méthodes de traitement de la non-réponse non monotone existantes peuvent être décrites brièvement comme il suit. L'approche paramétrique suppose des modèles paramétriques pour la propension à répondre ainsi que pour la distribution conjointe de la variable étudiée sur les divers points dans le temps (par exemple, Troxel, Hartington et Lipsitz 1998, Troxel, Lipsitz et Hartington 1998). Cependant, la validité de l'approche paramétrique dépend de la spécification correcte des modèles paramétriques. Vansicek et Robins (2007) ont proposé des méthodes sous certains modèles de la propension à répondre au temps t conditionnellement aux données observées antérieurement. Xu, Shao, Palla et Wang (2008) ont dérivé une procédure d'imputation sous les hypothèses que i) la propension à répondre au temps t dépend uniquement des valeurs de la variable étudiée au temps $t - 1$ et ii) la variable étudiée à différents points dans le temps est une chaîne de Markov. Une autre approche, que nous appellerons censure, consiste à créer un ensemble de données présentant une « non-réponse monotone » en écartant toutes les valeurs observées de la variable étudiée auprès d'un sujet

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706. Courriel : shao@stat.wisc.edu ; Martin Klein, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C. 20233 ; Jing Xu, Department of Statistics, University of Wisconsin, Madison, WI 53706.

Thoumi, F. (2003). *Illegal Drugs, Economy, and Society in the Andes*. Woodrow Wilson Center Press, Washington, États-Unis.

Tortico, J., Pohlan, H. et Janssens, M. (2005). Alternatives for the transformation of drug production areas in the chapare region, Bolivia. *Journal of Food, Agriculture and Development*, 3, 3-4.

Tourangeau, R., et Yam, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 5, 859-883.

Zufferey, A., Michaud, P., Jeannin, A., Berchtold, A., Chossis, L., van Melle, G. et Suris, J. (2007). Cumulative risk factors for adolescent alcohol misuse and its perceived consequences among 16 to 20 year old adolescents in Switzerland. *Preventive Medicine*, 45, 2-3, 233-239.

Varkey, P., Balakrishna, P., Prasad, J., Abraham, S. et Joseph, A. (2000). The reality of unsafe abortion in a rural community in South India. *Reproductive Health Matters*, 8, 16, 83-91.

Willis, G. (2005). *Cognitive interviewing. A tool for improving questionnaire design*. Sage Publications, Inc. États-Unis.

- Caballero, V., Dietz, E., Tabada, C. et Anduaga, J. (1998). Diagnóstico Rural Participativo de las Cuenclas Alto Inambari y Alto Tambopeata Provincia de Sandia, Departamento de Puno. GTZ, Lima, Pérou.
- Caulkins, J., Reuter, P., Iquchi, M. et Chiesà, J. (2005). How goes the War on Drugs? An Assessment of U.S. Drug Problems and Policy. RAND Drug Policy Research Center. Etats-Unis.
- Collins, J. (1984). The maintenance of peasant coffee production in a peruvian valley. *American Ethnologist*, 11, 3, 413-438.
- Commission des supérieurs (2005). Alternative Development: A Global Thematic Evaluation. Rapport de synthèse final. Quarante-huitième session E/CN.7/2005/CRP.3. Autriche.
- Coutts, E., et Jann, B. (2008). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). ETH Zurich Sociology. Document de travail, 3.
- Davalos, L., Bejarano, A. et Correa, L. (2008). Disabusing cocaine: pervasive myths and enduring realities of a globalised commodity. *International Journal of Drug Policy*, 20, 5, 381-386.
- Davis, C., Thake, J. et Vilhena, N. (2009). Social Desirability Biases in Self-Reported Alcohol Consumption and Harms. Addictive Behaviors. Article sous presse.
- Durand, F. (2005). El Problema Cocacero y el Comercio Informal para Uso Tradicional. Debate Agrario 39. Lima, Pérou.
- Département d'Etat des Etats-Unis (2009). International Narcotics Control Strategy Report. Volume I: Drug and Chemical Control. Bureau for International Narcotics and Law Enforcement Affairs. Etats-Unis.
- Fergusson, D., Boden, J. et Horwood, L. (2008). The developmental antecedents of illicit drug use: Evidence from a 25-Year longitudinal study. *Drug and Alcohol Dependence*, 96, 165-177.
- Fu, H., Darroch, J., Henshaw, S. et Kolb, E. (1998). Measuring the extent of abortion underreporting in the 1995 National Survey of Family Growth. *Family Planning Perspectives*, 30, 3, 128-138.
- García, J., et Antezana, J. (2009). Diagnóstico de la Situación del Desvío de IQ al Narcotráfico. ConsultAndes and DEVIDA. Lima, Pérou.
- Gibson, B., et Godoy, R. (1993). Alternatives to coca production in Bolivia: A computable general equilibrium approach. *World Development*, 21, 6, 1007-1021.
- Hernan, A. (1990). Tradición y represión: Dos experiencias en america del sur. Dans *Coca, Coca y Narcotráfico. Laberinto en los Andes*, (Eds., García – D. Sayan), Comision Andina de Justias. Lima, Pérou.
- Holmstedt, B., Jaamaa, E., Leander, K. et Plowman, T. (1977). Determination of cocaine in some South American species of erythroxylum using mass fragmentography. *Phytochemistry*, 16, 1753-1755.
- Statistique Canada, N° 12-001-X au catalogue
- Ibanez, M. et Carlsson, F. (2010). A survey-based choice experiment on coca cultivation. *Journal of Development Economics*, 93, 2, 249-263.
- INEI (2007). Censos Nacionales 2007: XI de Población y VI de Vivienda. Lima, Pérou.
- Johnson-Hanks, J. (2002). The lesser shame: Abortion among educated women in southern Cameroon. *Social Science & Medicine*, 55, 8, 1337-1349.
- Mansfield, D. (2006). Development in Drug Environment: A Strategic Approach to Alternative Development. Article de discussion. Development Oriented Drug Control Program. GTZ, Allemagne.
- Mensch, B., Hewett, P. et Erulkar, A. (2003). The reporting of sensitive behavior by adolescents: A methodological experiment in Kenya. *Demography*, 40, 2, 247-268.
- Obando, E. (2006). U.S. Policy toward Peru: At odds for twenty years. Dans *Addicted to Failure. U.S. Security Policy in Latin America and the Andean Region*, (Eds., B. Loveman), Rowman & Littlefield Publishers Inc. Etats-Unis.
- Office of Technology Assessment (1993). Alternative Coca Reduction Strategies in the Andean Region. U.S. Congress. OTA-F-556. Washington, Etats-Unis.
- OICS (2009). Report on the International Narcotics Control Board for 2009. United Nations Publication. New York, Etats-Unis.
- ONUDC (2001). Alternative Development in the Andean Area. The UNDCP Experience. Edition révisée. ODCPC Studies on Drugs and Crime. New York, Etats-Unis.
- ONUDC (2009). Peru. Monitoreo de Cultivos de Coca 2008. Lima, Pérou.
- ONUDC (2011). Peru. Monitoreo de Cultivos de Coca 2010. Lima, Pérou.
- Punch, K. (2003). Survey research. The basics. *Sage Publications, Inc.* Royaume-Uni.
- Rivera, M., Audechide, A., Cartmell, L., Torres, C. et Langsjoen, O. (2005). Antiquity of coca – Leaf chewing in the south central Andes: A 3000 year archaeological record of coca – Leaf chewing from Northern Chile. *Journal of Psychoactive Drugs*, 37, 4, 455-458.
- Rospigliosi, F. (2004). Analisis de la Encuesta DEVIDA-INEL. Dans *El Consumo Tradicional de la Hoja de Coca en el Peru*, (Ed., F. Rospigliosi). Instituto de Estudios Peruanos. Lima, Pérou.
- Singer, E., Hippler, H. et Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research*, 4, 3.
- Strutin, L. (2001). Assessing alcohol consumption: developments from qualitative research methods. *Social Science & Medicine*, 53, 2, 215-226.

Annexe 2

Statistiques descriptives comparatives entre toutes les observations et les non-répondants à la question de nature délicate

Variables	Toutes les observations*	Non-répondants à la question de nature délicate
Âge	42,2 (12,6)	45,9 (9,9)
Hommes (%)	94,3***	100***
Aymara (%)	81,8**	94,7**
Nombre d'enfants	3,0**	4,1**
Années de scolarité	8,4 (2,0)	7,5 (2,9)
Superficie totale (hectares)	7,9 (8,3)	6,8 (3,2)
Superficie consacrée au café (hectares)	2,2 (1,8)	2,5 (1,2)
Superficie de forêt secondaire (zone en jachère)	1,6 (2,3)	1,4 (1,1)
Superficie de forêt primaire (hectares)	4,0 (7,3)	2,9 (3,3)
Superficie consacrée aux aliments de base (hectares)	0,5 (0,7)	0,6 (0,6)
Aucune autre activité économique (%)	47,5	57,9
Aversion élevée au risque (%)	28,6***	73,7***
Importance de respecter les lois nationales (%)	84,3	89,5
Diminution de la confiance au cours des cinq dernières années (%)	16,8	26,3
Participation à des activités communautaires en 2007 (%)	89,4	89,5
Agriculteurs masquant la coca (%)	67,7	73,7
Agriculteurs utilisant la coca comme médicament (%)	72,0*	84,2*
Perception qu'il est facile de vendre les feuilles de coca (%)	23,6	27,8
Nombre d'observations	477	19

Les écarts-types sont entre parenthèses pour les variables continues.

a) Toutes les observations sans les non-répondants à la question de nature délicate.
Les moyennes pour les non-répondants sont statistiquement différentes de l'ensemble de l'échantillon (test T avec variances inégales) au :
* niveau de signification de 0,1 ; ** niveau de signification de 0,05 ; *** niveau de signification de 0,01.

Source : Calculs de l'auteur.

Bibliographie

- Allen, C. (1981). To be Quechua: The symbolism of coca chewing in highland Peru. *American Ethnologist*, 8, 1, 157-171.
- Barnett, J. (1998). Sensitive questions and response effects: An evaluation. *Journal of Managerial Psychology*, 13, 1/2, 63-67.
- Bedoya, E. (2003). Estrategias productivas y el riesgo entre los cocaleiros del valle de los ríos apurímac y gené. Dans *Amazonia: Procesos Demográficos y Ambientales*, (Eds., C. Aramburu et E. Bedoya), Consorcio de Investigación Económica y Social. Lima, Pérou.
- Bureau du Pérou de l'ONUDC (1999). Desarrollo Alternativo del Inambari y Tambopata. Documento de Proyecto A/D/PER/99/D/96. Disponible au : <http://www.onudc.org.pe/web/Html/Templates/proyectos.htm> (accessible le 15 juin 2009).
- Binswanger, H. (1980). Attitude towards risk: Experimental measurement in rural India. *American Journal of Economics*, 62, 395-407.
- Boivin, G., Griffin, K., Diaz, T., Scheier, L., Williams, C. et Epstein, J. (2000). Preventing illicit drug use in adolescents: Long-term follow-up data from a randomized control trial of a school population. *Addictive Behaviors*, 25, 5, 769-774.

Les producteurs commerciaux de coca peuvent profiter de cette situation et continuer de cultiver de la coca en prétextant des usages traditionnels.

Remerciements

La recherche a été financée par le BMZ (Ministère fédéral de la coopération économique et du développement, Allemagne) et le DAD (Service d'échanges universitaires de l'Allemagne), ainsi que par le LACBEP (Programme latino-américain et caribéen d'économie environnementale).

Annexe 1

Sections pertinentes du questionnaire

A) Présentation :

Bonjour/bonsoir. Je m'appelle _____ et j'étudie à _____. Nous menons une enquête pour déterminer les risques et les vulnérabilités auxquels sont exposés les producteurs de café dans votre collectivité. Les directeurs des coopératives de café connaissent cette enquête et croient que les résultats pourraient profiter à la collectivité. Si vous décidez de répondre à notre questionnaire, vous pouvez sauter des questions ou vous retirer de l'étude à tout moment. Les données recueillies dans le cadre de cette enquête demeureront CONFIDENTIELLES et serviront uniquement à des fins UNIVERSITAIRES. Vos réponses et vos opinions sont extrêmement importantes pour la coopérative et pour nous. Êtes-vous prêt à répondre à certaines questions ?

a) Oui (poursuivre)
b) Non (remercier le répondant, interrompre l'enquête et indiquer les caractéristiques de la personne dans la présentation 1)

B) Questions relatives à la coca :

Dans cette partie de l'enquête, nous vous posons des questions concernant les usages et la culture de la coca. Veuillez vous rappeler que cette enquête est anonyme et qu'il n'y a pas de réponses correctes ou incorrectes.

Masquez-vous des feuilles de coca ?
a) Oui
b) Non

Utilisez-vous les feuilles de coca à des fins médicinales ?
a) Oui
b) Non

Vous croyez-vous obligé d'offrir des feuilles de coca à vos invités pendant les activités d'ayni et de minka ?
a) Oui
b) Non

Utilisez-vous les feuilles de coca pour le paiement de travailleurs de l'extérieur ?
a) Oui
b) Non

Utilisez-vous des feuilles de coca comme produit d'échange ou comme cadeau pour des amis ou parents ?
a) Oui
b) Non

Combien de petits arbustes de coca y a-t-il sur votre parcelle agricole ?
a) Oui
b) Non

C) Question sur l'aversion au risque :

Ceci est un jeu. Avant de jouer, vous devez choisir l'une des options affichées ci-dessous. Puis, vous devez tirer à pile ou face. Si, par exemple, vous avez choisi l'option H et que je tire à pile ou face et que la pièce tombe sur face, vous ne gagnez pas d'argent du tout. Toutefois, si le résultat est pile, vous gagnez 200 soles. Par ailleurs, si vous avez choisi l'option A, vous recevrez 50 soles, peu importe si le résultat est pile ou face. Laquelle des options parmi celles qui précèdent choisissez-vous avant que je lance la pièce ?

OPTION		
Si le résultat est face, vous gagnez :		
A	50 soles	Si le résultat est pile, vous gagnez :
B	45 soles	
C	40 soles	
D	35 soles	
E	30 soles	
F	20 soles	
G	10 soles	
H	0 sol	
	50 soles	

Notre échantillon de 477 répondants (en excluant les agriculteurs qui n'ont pas indiqué leur numéro d'enregistrement à la coopérative et les non-répondants à la question de nature délicate) ont déclaré au total 960 000 arbuscules de coca. Cet échantillon représente 14,6 % du total des 3 265 membres de coopératives à SPP. Ainsi, si l'on extrapole en fonction du nombre total de membres de coopératives situées dans le district de SPP, on obtient un total de 6,6 millions d'arbuscules de coca. Par ailleurs, nous devons tenir compte du fait que la vallée du cours supérieur du Tambopata comprend aussi le district de San Juan del Oro, qui compte à peu près la même population que le district de SPP (INEI 2007). Selon l'hypothèse très solide que les agriculteurs de SPP se comportent de la même façon que les agriculteurs de San Juan del Oro, à tout le moins du point de vue de la culture de la coca, cela ferait doubler le nombre d'arbuscules de coca pour l'ensemble de la vallée du cours supérieur du Tambopata, ce nombre atteignant environ 13,2 millions. Cette dernière estimation se situe entre 35 % et 40 % des 32,9 à 37,6 millions obtenus à partir des données recueillies par satellite de l'ONUDC. Elle se situe dans la fourchette attendue de déclaration pour les questions de nature délicate. Dans le cas des questions sur l'avortement, cette fourchette se situe entre 35 % et 59 % (Fu, Darroch, Henshaw et Kolb 1998), et pour l'utilisation d'opiacés ou de la cocaïne, entre 30 % et 70 % (Tourangeau et Yan 2007).

4. Sommaire et conclusions

La coca, qui est la matière première pour la production de la cocaïne, est cultivée en Colombie, au Pérou et en Bolivie. Dans ces deux derniers pays, les usages traditionnels de la coca par les populations indigènes remontent à environ 3000 ans avant Jésus-Christ (Rivera et coll. 2005). Néanmoins, les questions aux agriculteurs sur l'étendue de leur culture de coca sont considérées comme délicates. Les producteurs de coca craignent les programmes d'éradication, même s'ils ne vendent pas de coca pour le narcotrafic, parce qu'il est difficile de faire une distinction entre les producteurs de coca à des fins commerciales, et ceux qui produisent pour leur propre consommation. Ainsi, les agriculteurs ont tendance à ne pas participer aux enquêtes, à ne pas répondre aux questions de nature délicate ou à sous-déclarer la superficie de leurs aires de culture de la coca, afin de réduire les possibilités d'identification, en vue d'une éradication possible.

Dans ce contexte, les procédures de collecte des données au niveau des ménages doivent être évaluées, ainsi que les stratégies pour réduire la non-réponse et la déclaration incorrecte. La plupart des stratégies utilisées dans notre zone de recherche au Pérou ont été fondées sur les pratiques exemplaires comprises dans des ouvrages publiés. Parmi les stratégies qui ont fonctionné dans notre cas figurent l'établissement d'un lien de confiance avec les agriculteurs au moyen d'une lettre de présentation d'un directeur d'une

La validité des réponses individuelles des agriculteurs concernant l'étendue de leur culture de coca ne peut être vérifiée, parce que peu de recherches empiriques ont été effectuées à ce sujet, et qu'il n'y a pas d'autres sources au niveau des ménages pour confirmer ces données. Ainsi, la portée de la déclaration incorrecte a été évaluée à partir de données agrégées. Les résultats laissent supposer que les agriculteurs n'ont déclaré que de 35 % à 40 % de leur superficie réelle de culture de la coca. Toutefois, ces valeurs se situent à l'intérieur des fourchettes attendues pour les réponses aux questions de nature délicate. Du point de vue de la non-réponse à l'enquête et de la non-réponse à la question de nature délicate, les résultats étaient plus encourageants, indiquant des valeurs de 10 % et d'environ 4 %, respectivement.

À un moment de la tenue de l'enquête, nous avons principalement profité des célébrations et des assemblées générales des coopératives pour lesquelles les agriculteurs se réunissent dans le village, ceux-ci étant autrement très disséminés dans la forêt tropicale humide. L'enquête a suivi une méthode d'échantillonnage de commodité, mais il a été possible de vérifier la représentativité de l'échantillon parce que tous les agriculteurs sont enregistrés dans l'une des coopératives de la zone visée par la recherche. L'échantillon obtenu a été comparé à un échantillon aléatoire simple simulé sans remise, dans lequel chaque agriculteur avait la même probabilité d'être sélectionné au hasard à partir des listes de membres des coopératives. Il n'y avait pas de différences significatives dans les fonctions de distribution, ce qui fait que l'échantillon est équivalent à un échantillon aléatoire simple. Le principal inconvénient de cette approche est, qu'après l'interview, nous avons dû demander aux répondants leur numéro de membre de coopérative. Même si on a dit au répondant que ce numéro n'était pas joint à leur questionnaire, certains agriculteurs peuvent avoir eu des doutes à ce sujet et cela pourrait avoir eu des effets sur la crédibilité de l'assurance de la confidentialité dans les interviews suivantes, les agriculteurs s'étant passé le mot.

Par ailleurs, la comparaison des caractéristiques des non-répondants aux questions de nature délicate et du reste des non-répondants montre que les non-répondants sont très réfractaires au risque. Même si le nombre de non-répondants était faible (moins de 4 % de l'échantillon total), cela pourrait laisser supposer que la principale raison de la non-réponse partielle est la crainte des conséquences de la transmission des données à des tiers.

Il existe un rapport statistiquement significatif entre la culture de la coca et les usages traditionnels. Un pourcentage plus élevé de producteurs de coca que de non-producteurs masquent la coca et l'utilisent comme médicament. Qui plus est, un plus grand nombre de producteurs de coca trouvent plus facile de vendre les feuilles de coca que les non-producteurs, dans le cas hypothétique où ils cultiveraient la coca à des fins commerciales.

Enfin, il est important de mentionner que le nombre moyen d'arbustes de coca est relativement faible, ce qui pourrait être attribuable à une sous-déclaration des zones de culture commerciale de la coca ou à la culture de la coca pour consommation propre seulement, ou les deux. Il n'est pas possible de faire de distinction entre ces deux scénarios, ce qui fait qu'il est plus facile pour les producteurs commerciaux de coca de se faire passer pour des producteurs de coca pour des usages traditionnels.

3.2 Problèmes de validation

On ne peut pas vérifier directement la validité des réponses individuelles parce qu'il existe peu de recherches empiriques antérieures sur ce sujet et pas d'autres sources

Tableau 3
Statistiques descriptives comparatives entre les producteurs de coca et les non-producteurs

Variable	Producteurs de coca	Non-producteurs de coca
Âge	42,5	41,7
Hommes (%)	(12,7)	(12,5)
Aymara (%)	81,4	82,5
Nombre d'enfants	3,0	2,9
Années de scolarité	8,2*	8,7*
Superficie totale (hectares)	(3,3)	(3,3)
Superficie consacrée au café (hectares)	2,2	2,2
Superficie de forêt secondaire (zone en jachère)	(2,0)	(1,4)
Superficie de forêt primaire (hectares)	(7,5)	(7,0)
Superficie consacrée aux aliments de base (hectares)	0,5	0,5
Aucune autre activité économique (%)	(0,7)	(0,6)
Aversion élevée au risque (%)	46,8	48,9
Importance de respecter les lois nationales (%)	81,9**	88,6**
Diminution de la confiance au cours des cinq dernières années (%)	19,3**	12,5**
Participation à des activités communautaires en 2007 (%)	92,0**	84,7**
Agriculteurs masquant la coca (%)	76,0***	53,1***
Agriculteurs utilisant la coca comme médicament (%)	81,7***	54,8***
Perception qu'il est facile de vendre les feuilles de coca (%)	26,4**	18,5**
Nombre d'arbustes de coca	3 093	-
Nombre d'observations	(6 710)	305
172		

Les écarts-types sont entre parenthèses pour les variables continues.
Les moyennes pour les producteurs de coca et les non-producteurs de coca sont statistiquement différentes (test T avec variances inégales) au :
* niveau de signification de 0,1 ; ** niveau de signification de 0,05 ; *** niveau de signification de 0,01.
Source : Calculs de l'auteur.

3. Résultats de l'enquête et problèmes de validation

3.1 Résultats de l'enquête

Le taux de réponse à l'enquête se situe à environ 90 %, ce qui est bien au-dessus du taux de réponse minimum recommandé de 60 % (Punch 2003). Dans les 496 questionnaires remplis, 19 répondants (moins de 4 %) n'ont pas répondu aux questions liées à la coca. Lorsque l'on compare les statistiques descriptives des variables socioéconomiques, institutionnelles et liées à la coca, on note certaines différences significatives entre toutes les observations (sans les non-répondants) et les « non-répondants à la question de nature délicate » (voir l'annexe 2). Les non-répondants à la question de nature délicate étaient tous de sexe masculin, avec un pourcentage plus grand de descendance ethnique Aymara, et un plus grand nombre d'enfants. En outre, un pourcentage plus élevé d'entre eux utilisait la coca à des fins médicinales. Il est intéressant de constater qu'un nombre significativement plus élevé de non-répondants sont très réfractaires au risque (73,7 %), comparativement à tous les autres répondants (28,6 %). Cela pourrait indiquer une crainte possible de la part des « non-répondants à la question de nature délicate » que les intervieweurs divulguent les renseignements à des tiers. Le contexte du test d'aveu au risque suivi par Binswanger (1980) est présenté à l'annexe 1c.

Des statistiques descriptives comparatives de base des producteurs de coca et des non-producteurs sont présentées dans le tableau 3. Le nombre de questionnaires valides était de 477, si nous ne tenons pas compte des non-répondants à la question de nature délicate. Parmi eux, 64 % ont indiqué qu'il y a pas de différences statistiquement significatives en ce qui a trait aux caractéristiques socioéconomiques générales (âge, sexe, groupe ethnique, et nombre d'enfants) entre les producteurs de coca et les non-producteurs. La seule différence a été observée au chapitre de la scolarité. Les non-producteurs comptent plus d'années de scolarité que les producteurs. Chez les producteurs de coca, on retrouve une moins grande superficie de forêt au total et une moins grande superficie de forêt primaire, et davantage de terres en jachère que chez les non-producteurs, même si ces différences ne sont pas statistiquement significatives. Chez les producteurs de coca et les non-producteurs, on retrouve des superficies consacrées à la production de café et d'aliments comparables. Par contre, les producteurs de coca et les non-producteurs affichaient des différences statistiques significatives dans les variables du capital social. Un plus grand nombre de non-producteurs que de producteurs trouvent important de respecter les lois nationales. Par ailleurs, un moins grand nombre de non-producteurs que de producteurs ont vu leur confiance à l'égard de leur voisin diminuer au cours des cinq dernières années et ont collaboré à des activités communautaires au cours de la dernière année.

d'enregistrement à la coopérative était manquant. Dans deux cas, les agriculteurs ont refusé de fournir ces renseignements et, dans dix cas, les intervieweurs ont oublié de demander aux répondants leur numéro d'enregistrement à la fin de l'interview. Par conséquent, l'absence de ces renseignements a davantage été associée à une erreur de l'intervieweur qu'au refus de l'agriculteur de fournir ces renseignements.

Tableau 2
Nombre de répondants par coopérative

Nombre total	Taille de l'échantillon	Pourcentage des membres	SPP des coopératives à l'enquête	coopératives des interviewés	(%)
Coopérative 1	756	106			
Coopérative 2	911	138			
Coopérative 3	887	138			
Coopérative 4	711	114			
Total	3 265	496			

Source : Enquête de l'auteur.

Afin de vérifier la représentativité de l'échantillon, la répartition des numéros d'enregistrement aux coopératives obtenus auprès de l'échantillon de l'enquête a été comparée à la répartition des numéros d'enregistrement aux coopératives d'un échantillon aléatoire simple simulé, sans remise, tiré des listes des coopératives. Les listes des coopératives ont été classées par numéro d'enregistrement des membres, et les numéros d'enregistrement ont été associés à la date d'enregistrement des membres. Ainsi, la plupart des agriculteurs plus âgés ont des numéros d'enregistrement plus bas, et les agriculteurs plus jeunes, des numéros plus élevés. Malheureusement, les coopératives n'avaient pas d'autres renseignements sur les membres, comme la superficie totale des terres, ou encore les hectares consacrés au café ou à la coca, pouvant servir à sélectionner un échantillon aléatoire stratifié. Deux types de tests ont été utilisés pour la comparaison des échantillons : un test de la somme des rangs pour deux échantillons de Wilcoxon (Mann-Whitney) et un test pour l'égalité des fonctions de distribution pour deux échantillons de Kolmogorov-Smirnov. Le premier test sert à déterminer dans quelle mesure il est probable que les deux groupes proviennent de la même distribution, et repose sur le principe que les différences observées sont causées par une fluctuation du hasard. Le deuxième test est similaire au premier, mais il est aussi sensible aux différences dans l'empilement et la forme des fonctions de distribution cumulative empirique des deux groupes. Les résultats des deux tests n'ont pas rejeté l'hypothèse nulle de l'égalité de la distribution entre l'échantillon de l'enquête et l'échantillon aléatoire simple simulé, au niveau de signification de 0,05. Ainsi, les résultats montrent que l'échantillon de l'enquête est équivalent à un échantillon aléatoire simple et, par conséquent, est représentatif de la population à l'étude.

les Aymaras, qui utilisent couramment des diminutifs dans leurs conversations de tous les jours. Par ailleurs, le format de cette question pouvait impliquer indirectement que l'intervé(e)ur s'attendait à ce que le répondant ait un petit nombre d'arbustes de coca, ce qui aurait probablement donné lieu à une sous-déclaration. Par conséquent, même si des non-réponses ont été évitées grâce à cette dernière question, on s'attendait à une certaine sous-déclaration.

Période de tenue de l'enquête et contexte de collecte des données

Les parcelles agricoles des agriculteurs sont disséminées dans la forêt tropicale humide montagneuse amazonienne au Pérou. Il était difficile de joindre les agriculteurs sur leurs terres pendant l'enquête. Par conséquent, pour mener l'enquête, nous avons principalement profité de la célébration de la fête de Saint-Pierre et des réunions de l'assemblée générale des coopératives, en juin et août 2008, respectivement, qui sont une occasion de rencontre des agriculteurs sur la place du village. La participation aux réunions de l'assemblée générale est obligatoire pour tous les membres des coopératives, ce qui fait que les répondants ciblés étaient présents à ces activités. La seule façon d'entrer sur la place du village ou d'en sortir est par une route non pavée. Afin de profiter de cette situation, l'enquête a été menée à partir d'une grande tente érigée sur la route non pavée ces journées-là. La tente comprenait dix cubicules, un pour chaque paire d'intervé(e)urs et de répondants. On n'a pas assuré une protection absolue de la vie privée, parce que pendant l'étude pilote, on a déterminé que les agriculteurs n'étaient pas à l'aise d'être les « seuls » à être interviewés ; ils préféraient en voir d'autres être interviewés en même temps qu'eux. Toutefois, les agriculteurs ne pouvaient pas entendre les réponses des autres. Comme tous les agriculteurs devaient utiliser la même route non pavée pour se rendre sur la place du village, peu importe leur provenance géographique, les biais géographiques potentiels qui, quant à eux, peuvent être liés à des variables importantes, comme la taille de la ferme et le revenu, ont probablement été

Représentativité de l'échantillon

Une méthode d'échantillonnage de commodité a été utilisée, mais à la fin de l'enquête, nous avons demandé aux agriculteurs leur numéro d'enregistrement à la coopérative et nous avons utilisé les listes d'enregistrement pour intégrer la représentativité de l'échantillon. Le numéro d'enregistrement à la coopérative fourni par les agriculteurs était inscrit sur une feuille de papier distincte et n'était pas joint au questionnaire du répondant. Les répondants ont été informés de cette procédure et ont pu en être témoins. Les quatre coopératives visées par l'étude comptent 3 265 membres à SPPP. Le tableau 2 montre le nombre de répondants par coopérative. Le nombre de questionnaires remplis se chiffrait à 508. Au total, 12 répondants ont été exclus de l'échantillon parce que leur numéro

d'éviter les malentendus possibles et d'augmenter la confiance parmi les répondants. Au cours de l'étude pilote, certains agriculteurs ont indiqué des préoccupations concernant les programmes d'éradication de la coca financés à l'extérieur et, par conséquent, les références aux institutions externes ont été réduites au minimum. C'est donc dire que seuls des renseignements partiels ont été fournis aux répondants. Cela n'est pas courant, mais dans les circonstances particulières de l'étude, il n'y avait pas d'autre solution possible pour éviter les problèmes de sécurité potentiels.

En ce qui a trait à la formation, les intervieweurs ont un atelier de trois jours dans la zone visée par la recherche. Le même groupe d'intervé(e)urs a aussi mené l'étude pilote, afin de tester les questions et le questionnaire, avec comme objectif de déterminer les problèmes de compréhension, ainsi que de permettre la reformulation de l'élucidation, de mémorisation, de jugement et d'acceptabilité de la question de questions. L'étude pilote a aussi permis d'évaluer le rendement des intervieweurs et, dans certains cas, de déterminer les domaines nécessitant une formation sur mesure, selon la réaction concernant le rendement. Par exemple, au début, un des intervieweurs hésitait à poser les questions liées à la coca et cet intervieweur a obtenu un nombre plus élevé que la moyenne de non-réponses à la question de nature délicate. Après une formation adaptée, l'intervé(e)ur a pu modifier son approche d'interview.

Présentation de la question de nature délicate

La présentation de la question présupposait le caractère délicat du comportement à l'étude, comme l'ont montré Tourangeau et Yan (2007). Par conséquent, on n'a pas demandé en premier aux agriculteurs s'ils avaient des terres consacrées à la culture de la coca, puis la superficie totale de leur aires de culture de la coca, puis la superficie totale de leurs aires de culture de la coca (« Quelle est la superficie de votre aire de culture de la coca en mètres ou en hectares ? »). Toutefois, il a été déterminé pendant l'étude pilote que les agriculteurs n'étaient pas à l'aise avec cette question et la sautaient ou se retiraient simplement de l'enquête. Par conséquent, le format de la question a été changé et on a utilisé un libellé indulgent à la place. On a posé aux agriculteurs la question suivante : « Combien de « petits arbustes de coca » avez-vous sur votre terre agricole ? » Ainsi, l'agriculteur pouvait répondre : « Seulement quelques-uns, j'ai... ». Même si la différence est à peine perceptible, dans le premier cas, il était beaucoup plus difficile pour les agriculteurs de commencer leur réponse par « Seulement quelques-uns... ». Ainsi, leur réponse par « Seulement quelques-uns, j'ai... » était plus facile pour les agriculteurs d'ajouter des explications d'excuse à leur réponse, ce qui les a rendus plus détendus. Cette dernière présentation de la question de nature délicate avait aussi comme avantage d'utiliser de la terminologie que connaissent bien

d'alléger leurs soupçons. Un bref rappel de l'assurance de confidentialité a été inclus au milieu du questionnaire, avant les questions liées aux utilisations traditionnelles de la coca et avant la question de nature délicate sur la superficie consacrée à cette culture. Ce rappel se lisait comme suit : « Dans cette partie de l'enquête, nous vous posons des questions concernant les usages et la culture de la coca. Veuillez vous rappeler que cette enquête est anonyme et qu'il n'y a pas de réponses correctes ou incorrectes » (voir l'annexe 1b). Cela fait suite à Willis (2005), qui mentionne qu'il est important de présenter des questions d'entrée en matière et une annonce du passage au sujet délicat, afin de réduire les réticences des répondants.

Mode de collecte des données

On a initialement envisagé d'avoir recours à des questionnaires administrés avec papier et crayon pour la collecte des données, afin de réduire le biais lié aux intervieweurs. Toutefois, au cours de l'étude de faisabilité, il est devenu évident que de nombreux agriculteurs, même ceux qui avaient dépassé le niveau primaire d'études (52 % de la population ; INEI 2007) n'étaient pas capables de lire sans effort. Les agriculteurs travaillaient dans leurs champs presque toute la journée et n'ont pas beaucoup d'occasions de pratiquer leur habileté en lecture. De même, les interviewers audio-auto-administrés assistés par ordinateur (AAsAO), la méthode de prédilection pour la collecte de données sur des sujets délicats dans les pays développés (Mensch, Hewett et Eruikar 2003), dépassaient les limites de ce projet, en raison du manque d'équipement et d'entretien électrique, ainsi que de connaissances en informatique dans la zone visée par la recherche. Il est probable que l'utilisation d'ordinateurs aurait augmenté l'anxiété et les soupçons concernant l'enquête, comme l'ont décrit Mensch et coll. (2003) pour l'Afrique. Par conséquent, les interviewers sur place sont le mode de collecte qui a été sélectionné, et l'accent a été mis sur la sélection des intervieweurs, leur formation et leur comportement.

Sélection, formation et comportement des intervieweurs

Un des problèmes liés à la sélection des intervieweurs est l'absence de professionnels suffisamment scolarisés dans la zone visée par la recherche. Ainsi, un groupe de dix étudiants de l'université publique la plus proche, se trouvant à 16 heures de la zone visée par la recherche, ont été choisis comme intervieweurs. Tous les intervieweurs étaient de descendance Aymara ou Quechua ; on a tenté ainsi de faire correspondre partiellement les caractéristiques des intervieweurs et celles des répondants. On croyait que cela augmenterait la probabilité de participation, parce que l'appariement pouvait faire augmenter la confiance et la sympathie entre l'intervieweur et le répondant (Tourangeau et Yan 2007). Les intervieweurs se sont présentés comme des étudiants de l'université locale, et aucun renseignement supplémentaire n'a été fourni concernant une université ou une organisation à l'extérieur du pays finançant l'étude, afin

de ne pas susciter des soupçons. L'assurance de confidentialité a été incluse au milieu du questionnaire, avant les questions liées aux utilisations traditionnelles de la coca et avant la question de nature délicate sur la superficie consacrée à cette culture. Ce rappel se lisait comme suit : « Dans cette partie de l'enquête, nous vous posons des questions concernant les usages et la culture de la coca. Veuillez vous rappeler que cette enquête est anonyme et qu'il n'y a pas de réponses correctes ou incorrectes » (voir l'annexe 1b). Cela fait suite à Willis (2005), qui mentionne qu'il est important de présenter des questions d'entrée en matière et une annonce du passage au sujet délicat, afin de réduire les réticences des répondants.

Plusieurs stratégies peuvent contribuer à réduire les biais possibles liés à la nature délicate de la question, à la déclaration incomplète, à la non-réponse partielle et à la non-réponse d'unités, ainsi qu'à la sélection soignée du mode de collecte des données et du libellé de la question de nature délicate ; adaptation des caractéristiques et du comportement des intervieweurs (voir Coutts et Janm 2008 ; Tourangeau et Yan 2007). De plus amples renseignements sur la mise en œuvre de ces stratégies dans notre étude de cas figurent ci-après.

Établissement d'une relation de confiance et assurance de l'anonymat

Les agriculteurs des régions de culture de la coca ont tendance à ne pas faire confiance aux gens de l'extérieur. Dans cette région particulière, nous avons déterminé qu'ils font confiance aux directeurs des coopératives de café. Un des directeurs des coopératives de café a signé une lettre de présentation autorisant notre recherche sur la culture agricole. On a montré la lettre aux agriculteurs avant la tenue de l'enquête. Un essai pilote mené avec et sans la lettre de présentation a démontré que la lettre était importante pour réduire les refus de participation à l'enquête. Dans l'introduction à l'enquête, l'intervieweur a aussi indiqué que le directeur de la coopérative autorisait l'enquête parce qu'il s'attendait à ce que les résultats profitent aux membres. En outre, on a clairement dit aux agriculteurs, au début de l'enquête, que les données recueillies demeureraient confidentielles, et on a souligné l'objectif académique du questionnaire (voir l'annexe 1a). Cette assurance d'anonymat était courte et précise, afin de réduire les soupçons des agriculteurs, comme l'ont indiqué Singer, Hipplier et Schwarz (1992). La culture de la coca a été traitée comme un comportement courant et ordinaire dans la région visée par la recherche, et une assurance de confidentialité longue et élaborée aurait pu susciter des réserves chez les agriculteurs, plutôt que



Figure 1 Carte de la zone visée par la recherche

L'enquête finale comprenait un questionnaire structuré axé sur la production agricole et le capital social. Le questionnaire comptait 15 sections :

1. Renseignements généraux au sujet de l'agriculteur et du ménage
2. Renseignements généraux au sujet des terres agricoles et de la superficie réservée au café
3. Activités économiques additionnelles
4. Données sur la certification biologique
5. Capital social cognitif et identité
6. Information et communication
7. Aspirations personnelles et attitudes face au risque
8. Capital social structurel
9. Chocs covariés et idiosyncrasiques
10. Capital humain
11. Réseaux sociaux

Les éléments de l'enquête liés aux questions de nature délicate sont présentés à l'annexe I. La question sur la superficie de culture de la coca est une question de nature délicate pour les agriculteurs. Ceux qui cultivent des superficies importantes de coca craignent que les données fournies soient accessibles aux autorités responsables des programmes d'éradication. Ainsi, ils s'inquiètent parfois des conséquences possibles d'une réponse honnête, pour le cas où les données viendraient à la connaissance d'un tiers. Dans ces cas, l'anonymat doit être garanti aux agriculteurs. Ils peuvent aussi être tentés de fournir des réponses socialement désirables aux intervieweurs. La coca

Par conséquent, la recherche portant sur la coca est maintenant axée sur l'évaluation de la rentabilité de cette culture, par rapport à d'autres cultures commerciales (voir, par exemple, Gibson et Godoy 1993 ; Torricco, Pohlman et Janssens 2005). Différentes tentatives ont été faites pour remplacer la coca par d'autres cultures, mais il a été déterminé de façon générale que le remplacement de culture comme politique antidrogue a été un échec (ONUDDC 2001). Les décideurs et les chercheurs ont reconnu qu'il existe des déterminants socioéconomiques pertinents autres que la rentabilité économique pour justifier la culture de la coca. Il s'agit notamment du capital social (Thoumi 2003), ainsi que des fonctions d'épargne et de réserve financière pour les dépenses importantes (Bodoyá 2003 ; Mansfield 2006). Pour vérifier ces dernières hypothèses sur la superficie des aires de culture de la coca, on a besoin de bases d'exhaustives incluant des données sur les ménages.

La culture de la coca n'est pas illégale à proprement parler au Pérou (au cours des années 1990, le gouvernement péruvien avait comme principal objectif de « pacifier » le pays en combattant les groupes terroristes. Il a mis en œuvre ce que l'on appelle actuellement la « doctrine Fujimori ». L'hypothèse qui sous-tend cette doctrine est que la culture de la coca n'est pas de nature criminelle, mais plutôt attribuable à la pauvreté. Par conséquent, la doctrine Fujimori a fait en sorte de décriminaliser la culture de la coca, ce qui a diminué le besoin de protection des agriculteurs contre les associations terroristes et ce qui a par conséquent facilité la tâche du gouvernement dans sa lutte contre ces groupes violents (Obando 2006). Cela rend compte en partie de l'acceptation sociale des utilisations traditionnelles de la coca dans ce pays (ONUDDC 2001). Ainsi, le cadre juridique actuel semble faciliter le narcotrafic parce que la coca utilisée pour le commerce illégal peut être cultivée sous prétexte de son utilisation à des fins traditionnelles (OICS 2009 ; Durand 2005). Par conséquent, García et Antezana (2009) sont d'avis que certains agriculteurs vendent de la coca à des personnes qui semblent la négocier à des fins traditionnelles, mais qui sont plutôt des narcotrafiquants qui transforment les feuilles de coca à différents endroits, comme des petites villes à la frontière de la Bolivie.

Même si la culture de la coca n'est pas illégale, les régions où elle se pratique et qui sont perçues comme fournissant les narcotrafiquants (c'est-à-dire les régions ayant de grands champs de coca) peuvent être ciblées par le gouvernement pour la mise en œuvre de programmes d'éradication forcée (Obando 2006). L'éradication est susceptible d'entraîner de grandes pertes économiques pour les agriculteurs de coca, selon la superficie totale de leurs aires de culture. Ainsi, certains des agriculteurs peuvent être réticents à fournir des données sur le fait qu'ils cultivent ou non de la coca. On devrait aussi s'attendre à ce que certains des agriculteurs qui admettent cultiver de la coca ne déclarent pas la superficie totale de leurs aires de culture, parce qu'ils craignent que les grands champs de coca fassent l'objet d'une éradication.

Étant donné que la mesure exacte et fiable de la superficie des aires de culture de la coca suscite des préoccupations politiques et stratégiques, il est nécessaire pour les méthodologistes d'enquêter de déterminer comment enregistrer la déclaration honnête de données et la réponse à des questions de nature délicate concernant la culture de la coca. Le présent article suggère et évalue un certain nombre de stratégies, en vue d'augmenter à la fois la déclaration et la fiabilité des données au niveau des ménages dans une région éloignée de culture de la coca au Pérou.

Même si le sujet du présent article est particulièrement lié à la culture de la coca, les leçons apprises concernant la conception de l'enquête et sa mise en œuvre pourraient servir de référence dans le cadre d'autres sujets de nature délicate, comme les questions liées à la santé (par exemple, les mesures anticonceptionnelles et le comportement sexuel) ou des comportements indésirables (par exemple la consommation de drogue illégale) dans d'autres régions de différents pays.

L'article est structuré de la façon suivante : la section 2 décrit la collectivité du Pérou à l'étude, les stratégies particulières pour réduire la non-réponse et la déclaration incorrecte, ainsi que les leçons apprises de la collecte de données au moyen de questions de nature délicate dans la zone visée par la recherche. La section 3 présente les résultats de l'enquête liée à la culture de la coca et leur validation, tandis que la section 4 est constituée d'un sommaire des principaux résultats, suivis par la conclusion.

2. Collecte de données dans une collectivité cultivant la coca d'une région rurale du Pérou

2.1 Description de la zone visée par la recherche

La zone visée par la recherche est située dans la vallée du cours supérieur du Tambopata, à la frontière avec la Bolivie, l'une des zones les plus éloignées et difficiles d'accès de la forêt tropicale humide amazonienne au Pérou (Bureau du Pérou de l'ONUDDC 1999). Cette vallée est située dans le corridor de la conservation de la biodiversité de Vilcabamba-Amboro, à proximité des zones protégées nationales (voir la figure 1). L'ensemble de la population de la vallée du Tambopata est constituée d'immigrants, et plus particulièrement de descendants de la population indigène Aymara. Il s'agit d'un groupe ethnique autochtone originaire des régions des Andes et de l'Altiplano de l'Amérique du Sud. Au cours des années 1950, la plupart des agriculteurs étaient des immigrants saisonniers qui quittaient leurs terres de subsistance de l'Altiplano pendant trois à six mois par année et parcouraient les 320 km les séparant de la vallée du cours supérieur du Tambopata pour cultiver du

Collecte de données : expérience et leçons apprises au chapitre des questions de nature délicate dans une région éloignée de culture de la coca au Pérou

Jaqueline Garcia-Yi et Ulrike Grote¹

Résumé

La coca est une plante indigène de la forêt tropicale humide amazonienne, dont on extrait la cocaïne, un alcaloïde illégal. Les agriculteurs considèrent comme délicate les questions concernant la superficie de leurs aires de culture de la coca dans les régions éloignées où cette plante est cultivée au Pérou. Par conséquent, ils ont tendance à ne pas participer aux enquêtes, à ne pas répondre aux questions de nature délicate ou à sous-déclarer la superficie de leurs aires individuelles de culture de la coca. La mesure exacte et fiable des aires de culture de la coca est une source de préoccupations politiques et stratégiques, ce qui fait que les méthodologistes d'enquête doivent déterminer comment encourager la déclaration honnête de données et la réponse aux questions de nature délicate concernant la culture de la coca. Parmi les stratégies d'enquête appliquées dans notre étude de cas figuraient l'établissement d'un rapport de confiance avec les agriculteurs, l'assurance de la confidentialité, la correspondance entre les caractéristiques des intervieweurs et celles des répondants, la modification de la présentation des questions de nature délicate et l'absence d'isolement absolu des répondants au cours de l'enquête. Les résultats de l'enquête ont été validés au moyen de données recueillies par satellite. Ils semblent indiquer que les agriculteurs ont tendance à sous-déclarer la superficie de leurs aires de culture de la coca dans une proportion de 35 % à 40 %.

Mots clés : Coca ; cocaïne ; questions de nature délicate ; déclaration incorrecte ; non-réponse ; Pérou.

1. Introduction

Au cours des 30 dernières années, on a utilisé de plus en plus les enquêtes pour explorer les sujets délicats (Tourangeau et Yan 2007). Par exemple, on a utilisé des données d'enquête pour examiner les comportements « socialement indésirables », comme la prévalence de la consommation de drogue illégale (par exemple Botvin, Griffin, Diaz, Scheier, Williams et Epstein 2000 ; Ferguson, Boden et Horwood 2008), les avortements illégaux (par exemple Johnson-Hanks 2002 ; Varkey, Balakrishna, Prasad, Abraham et Joseph 2000) ou la consommation d'alcool chez les adolescents (par exemple Strunin 2001 ; Zufferey, Michaud, Jeannin, Bertschold, Chossis, van Melle et Surtis 2007). De telles enquêtes ont été couramment utilisées dans les recherches universitaires et l'analyse des politiques (Davis, Thake et Vilhena 2009), même si les questions de nature délicate ont toujours été perçues comme problématiques. Les réponses sont considérées comme sujettes aux erreurs et aux biais, parce que les répondants sous-déclarent constamment les comportements socialement indésirables (Barnett 1998 ; Tourangeau et Yan 2007). Les faibles taux de réponses présentent une préoccupation supplémentaire. Ceux qui sont sélectionnés pour une enquête peuvent simplement refuser d'y participer, ou ils peuvent y participer, mais refuser de répondre aux questions de nature délicate (Tourangeau et Yan 2007). Des enquêtes récentes au niveau des ménages ont intégré des questions de nature délicate concernant la superficie des

aires de culture de la coca (voir, par exemple, Ibanez et Carlsson 2010). La coca est un arbuste indigène de la forêt tropicale humide amazonienne en Amérique du Sud. On extrait la cocaïne de ses feuilles. La superficie des aires de culture de la coca représente 40 % en Colombie, 40 % au Pérou et 20 % en Bolivie de la superficie totale des aires de culture de la coca à l'échelle mondiale, soit 154 100 hectares (ONU/DC 2011). Au Pérou et en Bolivie, les feuilles de cette plante sont utilisées traditionnellement à de nombreuses fins, depuis environ 3000 ans avant Jésus-Christ (Riviera, Auerderheide, Cartmell, Torres et Langsjøen 2005) jusqu'à nos jours. Ces utilisations traditionnelles comprenaient principalement la mastication de la feuille de coca et l'absorption de tisane de feuilles de coca pour surmonter la fatigue, la faim et la soif, ainsi que pour soulager les symptômes du « mal de l'altitude » et les maux d'estomac, respectivement (Rospighiosi 2004). Depuis les années 1970, toutefois, la culture de la coca a atteint des sommets, en raison de son utilisation comme matière première pour la production de cocaïne (Caulkins, Reuter, Iguchi et Chiesa 2005). Le contenu en cocaïne de la feuille de coca est inférieur à 1 % et va de 0,13 % à 0,86 % (Holmstedt, Jaamaa, Leander et Plovman 1977). Par conséquent, les trafiquants de narcotiques ont besoin de grandes quantités de feuilles de coca pour obtenir suffisamment d'alcaloïde pour sa commercialisation sur le marché illégal. En général, la culture de la coca pour le narcotrafic est une activité rentable. En fait, le revenu des agriculteurs qui cultivent de la coca est supérieur de 54 % à celui de ceux qui n'en cultivent pas (Davalos, Begarano et Correa 2008).

- Lyberg : La qualité des enquêtes
- U.S. Office of Management and Budget (2002). *Guidelines for ensuring, and maximizing the quality, objectivity, utility, and integrity of information disseminated by Federal agencies*. Federal register, 67, 36, 22 février.
- U.S. Office of Management and Budget (2006a). *Standards and Guidelines for Statistical Surveys*. U.S. Office for Management and Budget.
- U.S. Office of Management and Budget (2006b). *Questions and answers when designing surveys for information collection*. U.S. Office for management and Budget.
- Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 2, 119-137.
- Weisberg, H. (2005). *The Total Survey Error Approach*. The University of Chicago Press.
- Weisman, E., Balyozov, Z. et Venter, L. (2010). IMF's data quality assessment framework. Document présenté à la Conférence on Data Quality for International Organizations, Helsinki, 6 au 7 mai.
- West, B., et Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 5, 1004-1026.
- Williamack, D., Nichols, E. et Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. Dans *Survey Nonresponse*, (Eds., R. Groves, D. Dillman, J. Eltinge et R. Little), 213-228.
- Zarkovich, S. (1966). *Quality of Statistical Data*. Food and Agricultural Organization of the United Nations : Rome, Italie.
- U.S. Federal Committee on Statistical Methodology (2001). *Measuring and Reporting Sources of Errors in Surveys*. Statistical Policy, document de travail 31, Washington, DC : U.S. Office of Management and Budget.
- Schles, P., Joiner, B. et Streibel, B. (1996). *The Team Handbook*. Joiner Associates Inc.
- Shewhart, W.A. (1939). *Statistical Methods from the Viewpoint of Quality Control*. U.S. Department of Agriculture, Washington, DC, Etats-Unis.
- Smith, T. (2011). Report on the International Workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. NORC/University of Chicago.
- Spencer, B. (1985). Optimal data quality. *Journal of the American Statistical Association*, 80, 564-573.
- Statistique Canada (2002). Le cadre d'assurance de la qualité de Statistique Canada, N° au catalogue 12-586-X1E, Ottawa.
- Statistique Canada (2009). *Statistique Canada : Lignes directrices concernant la qualité*, cinquième édition, Ottawa.
- Statistics Netherlands (1997). A self assessment of the Department of Statistical Methods. Document de recherche N° 9747, Statistics Netherlands.
- Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.
- Trewin, D. (2001). Importance d'une culture de la qualité. *Recueil : Symposium 2001, La Qualité des Données d'un Organisme Statistique : Une Perspective Méthodologique*, Statistique Canada.
- U.S. Bureau of the Census (1974). *Standards for Discussion and Presentation of Errors in Data*. U.S. Department of Commerce, Bureau of the Census.
- U.S. Federal Committee on Statistical Methodology (2001). *Measuring and Reporting Sources of Errors in Surveys*. Statistical Policy, document de travail 31, Washington, DC : U.S. Office of Management and Budget.

- Kennickell, A., Mulrow, E. et Scheuren, F. (2009). Paradata or process modeling for inference. Document présenté à la Conférence on Modernization of Statistics Production, Stockholm, Suède.
- Kiear, A.N. (1897). The representative method of statistical surveys. *Kristiania Videnskaps-selskabets Skriffter: Historisk-filosofiske Klasse*, (en norvégien), 4, 37-56.
- Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.
- Kish, L. (1995). *The Hundred Years' Wars of Survey Sampling*. Centennial Representative Sampling, Rome.
- Koiz, S. (2005). Reflections on early history of official statistics and a modest proposal for global coordination. *Journal of Official Statistics*, 21, 2, 139-144.
- Kreuter, F., Couper, M. et Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Lyberg, L. (2002). Training of survey statisticians in government agencies-A review. Communication sollicitée présentée à la réunion des Joint Statistical Meetings, American Statistical Association, New York.
- Lyberg, L., Bergdahl, M., Blanc, M., Booleman, M., Grünwald, W., Haworth, M., Japce, L., Jones, L., Komer, T., Linden, H., Lundholm, G., Madaleno, M., Rademacher, W., Signore, M., Zijlha, M.J., Tzougas, I. et van Brakel, R. (2001). Summary report from the Leadership Group (LEG) on Quality. Eurostat.
- Lyberg, L., et Couper, M. (2005). The use of paradata in survey research. Communication sollicitée présentée à la réunion de l'Institut International de Statistique, Sydney.
- Lynn, P. (Ed.) (2003). *Quality Profile: British Household Panel Survey: Waves 1 to 10: 1991-2000*. Colchester : Institute for Social and Economic Research.
- Lynn, P. (2004). Editorial: Measuring and communicating survey quality. *Journal of the Royal Statistical Society, Series A*, 167.
- Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.
- Mitchie, M. (1993). Data quality: A quest for standard indicators. *Proceedings of the International Conference on Establishment Surveys*. American Statistical Association, 729-734.
- Moeller, R. (2005). *Brink's Modern Internal Auditing*. Sixième édition. New York : John Wiley & Sons, Inc.
- Morganstein, D., et Marker, D. (1997). Continuous quality improvement in statistical agencies. Dans *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York : John Wiley & Sons, Inc., 475-500.
- Neter, J., et Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 305, 18-55.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- Neyman, J. (1938). *Lectures and Conferences on Mathematical Statistics and Probability*. U.S. Department of Agriculture, Washington, DC.
- OCDE (2011). Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities. OCDE.
- O'Muircheartaigh, C. (1997). Measurement errors in surveys: A historical perspective. Dans *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York : John Wiley & Sons, Inc., 1-25.
- Organisation internationale de normalisation (2006). *Marker, Opinion and Social Research*. ISO Standard N° 2052.
- Philips, P., et Fricker, S. (2011). Quality measures. Note, Office of Survey Methods Research, U.S. Bureau of Labor Statistics.
- Pink, B., Borowik, J. et Lee, G. (2010). The case for an international statistical innovation program-Transforming national and international statistics systems. Document présenté au Collaborational Leaders Workshop, 19 au 23 avril, Sydney, Australie.
- Platiek, R., et Sæmål, C.-E. (2001). Can a statistician deliver? *Journal of Official Statistics*, 17, 1, 1-20 et la discussion, 21-27.
- Reedman, L., et Julien, C. (2010). Current and future applications of the generic statistical business process model at Statistics Canada. Document présenté au Q Conference, Helsinki.
- Rosén, B., et Elvers, E. (1999). Quality concept for official statistics. *Encyclopedia of Statistical Sciences*. New York : John Wiley & Sons, Inc., mise à jour, Volume 3, 621-629.
- Scheuren, F. (2001). Quelle est l'importance de l'exactitude? *Recueil: Symposium 2001, La Qualité des Données d'un Organisme Statistique: Une Perspective Methodologique*, Statistique Canada.
- Schilling, E., et Neubauer, D. (2009). *Acceptance Sampling in Quality Control*, 2^e éd. Chapman and Hall/CRC.
- Statistique Canada, N° 12-001-X au catalogue

- Ericson, W. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Séries B*, 195-233.
- European Foundation for Quality Management (1999). *The EFQM Excellence Model*. Van Haren.
- Eurostat (2009a). ESS Standard for Quality Reports. Eurostat.
- Eurostat (2009b). ESS handbook for Quality Reports. Eurostat.
- Eurostat (2011a). European statistics Code of Practice. Eurostat.
- Eurostat (2011b). Quality assurance framework (QAF). Eurostat.
- Fellegi, I. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- Fellegi, I. (1996). Characteristics of an effective statistical system. *Revue Internationale de Statistique*, 64, 2, 165-197.
- Felme, S., Lyberg, L. et Olsson, L. (1976). *Kvalitetskydd av data*. (Protecting Data Quality). Liber (en suédois).
- Fienberg, S., et Tanur, J. (1996). Reconsidering the fundamental conditions of Fisher and Neyman on experimentation and sampling. *Revue Internationale de Statistique*, 64, 237-253.
- Fisher, R. (1935). *The Design of Experiments*. New York : Hafner.
- Frankel, M., et King, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 1, 65-87.
- Gleason, E. (2011). Centralizing LAN services. Note, National Agricultural Statistics Service, U.S. Department of Agriculture.
- Groves, R. (1989). *Survey Errors and Survey Costs*. New York : John Wiley & Sons, Inc.
- Groves, R. (2011). The structure and activities of the U.S. Federal Statistical System: History and recurrent challenges. *The Annals of the American Academy of Political and Social Science*, 631, 163, Sage.
- Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls, W. et Waksberg, J. (Eds.) (1988). *Telephone Survey Methodology*. New York : John Wiley & Sons, Inc.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. et Tourangeau, R. (2009). *Survey Methodology*. Deuxième édition. New York : John Wiley & Sons, Inc.
- Groves, R., et Heeringa, S. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, A*, 169, 439-457.
- Groves, R., et Lyberg, L. (2010). Total survey error: Past, present and future. *Public Opinion Quarterly*, 74, 5, 849-879.
- Hansen, M., et Hurwitz, W. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 517-529.
- Hansen, M., Hurwitz, W. et Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 32^e Session, 38, Partie 2, 359-374.
- Hansen, M., Hurwitz, W. et Kalton, G. (2001). Quality profiles in U.S. Statistical Agencies. *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm 14 au 15 mai 2001, CD-ROM.
- Hansen, M., Hurwitz, W., Marks, E. et Mauldin, P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190.
- Hansen, M., Hurwitz, W. et Madow, W. (1953). *Sample Survey Methods and Theory*. Volumes 1 et II. New York : John Wiley & Sons, Inc.
- Hansen, M., Hurwitz, W., Marks, E. et Mauldin, P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190.
- Hansen, M., Hurwitz, W. et Pritsker, L. (1964). The estimation and interpretation of gross differences and simple response variance. Dans *Contributions to Statistics*, (Ed., C. Rao). Oxford : Pergamon Press, 111-136.
- Hansen, M., Hurwitz, W. et Pritsker, L. (1967). Standardization of procedures for the evaluation of data: Measurement errors and statistical standards in the Bureau of the Census. Document présenté à la réunion de l'Institut International de Statistique, Sydney.
- Hansen, M., et Steinberg, J. (1956). Control of errors in surveys. *Biometrics*, 462-474.
- Hansen, M., et Voigt, R. (1967). Program guidance through the evaluation of uses of official Statistics in the United States Bureau of the Census. Document présenté à la réunion de l'Institut International de Statistique, Sydney.
- Holt, T., et Jones, T. (1998). Quality work and conflicting policy objectives. *Proceedings of the 84^e DGINS Conference*, 28 au 29 mai, Stockholm, Suède. Eurostat.
- Jabine, T., King, K. et Petrom, R. (1990). Survey of Income and Program Participation (SIPP): Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.
- Joiner, B. (1994). *Generation Management*. McGraw-Hill.
- Julien, C., et Born, A. (2006). Quality management assessment at Statistics Canada. *Proceedings of the Q Conference*, Cardiff, Royaume-Uni.
- Julien, C., et Royce, D. (2007). Quality review of key indicators at Statistics Canada. *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*, 1113-1120.
- Juran, J.M. (1988). *Juran on Planning for Quality*. New York : Free Press.
- Juran, J.M. (1995). *A History of Managing for Quality*. ASQC Quality Press.
- Juran, J., et Gryna, F. (Eds.) (1988). *Juran's Quality Control Handbook*, 4^e édition. McGraw-Hill.
- Kalton, G. (2001). Quelle est l'importance de l'exactitude ? *Recueil : Symposium 2001. La Qualité des Données d'un Organisme Statistique : Une Perspective Méthodologique*, Statistique Canada.
- Kalton, G., Wingate, M., Krawchuk, S. et Levine, D. (2000). *Quality Profile for SASS Rounds 1-3: 1987-1995*. Washington, DC : U.S. Department of Education.
- Kasprzyk, D., et Kalton, G. (2001). Quality profiles in U.S. Statistical Agencies. *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm 14 au 15 mai 2001, CD-ROM.
- Kasprzyk, D., et Kalton, G. (2001). Quality profiles in U.S. Statistical Agencies. *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm 14 au 15 mai 2001, CD-ROM.

- Bailar, B., et Dalenius, T. (1969). Estimating the response variance components of the U.S. Bureau of the Census' Survey Model. *Sankhyā*, B, 341-360.
- Biemer, P. (20010). Overview of design issues: Total survey error. Dans *Handbook of Survey Research*, (Eds., P. Marsden et J. Wright). Deuxième édition. Emerald Group Publishing Limited.
- Biemer, P., et Lyberg, L. (2003). *Introduction to Survey Quality*. New York : John Wiley & Sons, Inc.
- Biemer, P., et Lyberg, L. (2012). Short course on Total Survey Error. The Joint Program in Survey Methodology (JPSM), 16 au 17 avril, Washington, DC.
- Biemer, P., Trewin, D., Japel, L., Bergdahl, H. et Pettersson, A. (2012). A tool for managing product quality. Document présenté à la conférence de Q2012, Athènes.
- Blyth, B. (2012). ISO 20252: Turning frameworks into best practice. Document présenté à la conférence de Q2012, Athènes.
- Bohata, M. (2011). Fit-for-purpose statistics for evidence based policy making. Note, Eurostat.
- Bowley, A.L. (1913). Working-class households in reading. *Journal of the Royal Statistical Society*, 76(7), 672-701.
- Box, G. (1990). Good quality costs less? How come? *Quality Engineering*, 3, 1, 85-90.
- Box, G., et Friends (2006). *Improving Almost Anything: Ideas and Essays*. New York : John Wiley & Sons, Inc.
- Brackstone, G. (1999). La gestion de la qualité des données dans un bureau de statistique. *Techniques d'enquête*, 25, 2, 159-171.
- Brackstone, G. (2001). Quelle est l'importance de l'exactitude? *Recueil : Symposium 2001. La Qualité des Données d'un Organisme Statistique : Une Perspective Méthodologique*, Statistique Canada.
- Breyfogle, F. (2003). *Implementing Six Sigma*. Deuxième édition. New York : John Wiley & Sons, Inc.
- Brooks, C., et Bailar, B. (1978). An error profile: Employment as measured by the Current Population Survey. Document de travail 3, Office of Management and Budget, Washington, DC.
- Chakrabarty, R., et Torres, G. (1996). American Housing Survey: A Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.
- Colledge, M., et March, M. (1993). Quality management: Development of a framework for a statistical agency. *Journal of Business and Economic Statistics*, 11, 157-165.
- Colledge, M., et March, M. (1997). Quality policies, standards, guidelines, and recommended practices. Dans *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz et D. Trewin). New York : John Wiley & Sons, Inc.
- Statistique Canada. Québec, Canada.
- Edging, J. (2011). Aggregate and systemic components of risk in total survey error models. Document présenté au ITSEW 2011, 193-242.
- Edwards, W., Lindman, H. et Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Eisner, P., Hahn, M. et Junker, C. (2008). User satisfaction surveys in Eurostat and in the European Statistical System. Document présenté à la conférence Q, Rome, Italie.
- Drucker, P. (1985). *Management*. Harper Colophone.
- Doyle, P., et Clark, C. (2001). Quality profiles and data users. Document présenté à l'International Conference on Quality in Official Statistics (Q), Stockholm.
- Doherty, K. (2010). How business architecture renewal is changing IT at Statistics Canada. Document présenté au Meeting on the Management of Statistical Information Systems. Daejeon, Corée du Sud, 26 au 29 avril.
- Dillman, D. (1996). Why innovation is difficult in government surveys (avec discussions). *Journal of Official Statistics*, 12, 2, 113-198.
- De Vries, W. (1999). Are we measuring up...? Questions on the performance of national systems. *Revue Internationale de Statistique*, 67, 1, 63-77.
- Deming, W.E., et Geoffrey, L. (1941). On sample inspection in the processing of census returns. *Journal of the American Statistical Association*, 36, 215, 351-360.
- Deming, E. (1986). *Out of the Crisis*. MIT.
- Deming, E. (1950). *Some Theory of Sampling*. New York : John Wiley & Sons, Inc.
- Deming, E. (1944). On errors in surveys. *American Sociological Review*, 9, 359-369.
- Dalenius, T. (1985b). Relevant official statistics. *Journal of Official Statistics*, 1(1), 21-33.
- Dalenius, T. (1985a). *Elements of Survey Sampling*. Swedish Agency for Research Cooperation with Developing Countries. Stockholm, Suède.
- Dalenius, T. (1967). Nonampling Errors in Census and Sample Surveys. Rapport N° 5 du projet de recherche Errors in Surveys. Stockholm University.
- Dalenius, T.E. (1968). Official statistics and their uses. *Revue de l'Institut International de Statistique*, 26(2), 121-140.
- Dalenius, T. (1969). Designing descriptive sample surveys. Dans *New Developments in Survey Sampling*, (Eds., N.L. Johnson et H. Smith). New York : John Wiley & Sons, Inc.
- Couper, M. (1998). Measuring Survey Quality in a CASIC Environment. Document présenté au Joint Statistical Meetings. American Statistical Association, Dallas, TX.
- Statistique Canada. N° 12-001-X au catalogue

qualité telles que l'actualité, la comparabilité et l'accessibilité doivent être prises en collaboration avec les utilisateurs principaux, tandis que le fournisseur du service est responsable d'offrir la meilleure exactitude possible, compte tenu des diverses contraintes.

Les discussions sur la qualité des enquêtes et l'adoption de stratégies de gestion de la qualité ont-elles abouti à de meilleures données ? Nous ne le savons pas. La qualité des enquêtes n'a pas été évaluée selon un mode avant-après. La tendance est à l'accroissement de la normalisation et de la centralisation, ce qui devrait s'avérer rentable, mais quand il s'agit de la qualité des données, certains indicateurs pointent dans la mauvaise direction. Par exemple, dans de nombreux pays, les taux de non-réponse augmentent et les propriétés des erreurs dues à la collecte en mode mixte, à la traduction du matériel d'enquête et à d'autres caractéristiques de conception ne sont pas entièrement comprises ou varient d'une culture à l'autre. Il n'existe pas de formule de conception, ce qui entraîne des prises de décisions boiteuses concernant les compromis et des difficultés à décider de l'intensité avec laquelle les contrôles de qualité doivent être appliqués. La quête de pratiques exemplaires persiste dans les organismes d'enquête, mais leur mise en œuvre est difficile et éparpillée. Un rehaussement généralisé du niveau des compétences s'impose manifestement. Un programme de perfectionnement international structuré à l'intention des fournisseurs de services est nécessaire, de même qu'une collaboration internationale systématique en vue de déterminer les meilleurs moyens de concevoir et de mettre en œuvre les enquêtes. Nous devons mieux servir les utilisateurs en leur fournissant des données dont l'erreur est faible. Nous pouvons pour cela combiner plus judicieusement nos connaissances de la statistique et des phénomènes cognitifs avec les principes de gestion de la qualité. La note vraiment positive est l'attitude extraordinairement favorable à l'amélioration de la qualité dont témoignent les organismes statistiques partout dans le monde.

Bibliographie

- Aitken, A., Hörgren, J., Jones, N., Lewis, D. et Zijlho, M. (2004). *Handbook on improving quality by analysis of process variables*. Office for National Statistics, Royaume-Uni.
- Anderson, R., Kasper, J. et Frankl, F. (1979). *Total Survey Error: Applications to Improve Health Surveys*. San Francisco : Jossey-Bass.
- Apley, L., Carnuthers, P., Lee, G., Oehm, D. et Yu, F. (2011). Industrialisation of statistical processes, methods and technologies. Document présenté à la réunion de l'Institut International de Statistique. Dublin.

évaluations. Même si ces dernières révèlent des problèmes, il est préférable que ce soit le fournisseur qui les découvre plutôt que des entités externes. La plupart des utilisateurs ne souhaitent pas participer aux discussions au sujet des erreurs et des compromis entre les divers types d'erreurs, et ce, pour de bonnes raisons. Elles sont simplement trop techniques et obscures. Si nous admettons qu'un processus de bonne qualité est une condition préalable à un produit de bonne qualité, nous devrions améliorer progressivement les processus afin qu'ils s'approchent de la situation parfaite d'absence de biais. De cette façon, la variance d'une estimation devient une bonne approximation de l'erreur quadratique moyenne.

Malgré des discussions sans fin et une myriade de projets d'amélioration de la qualité des enquêtes, les pratiques n'ont guère changé (Lynn 2004 ; Pink, Borowik et Lee 2010 ; Groves 2011 ; Bohata 2011). Le manque de compétence au sein des organismes d'enquête est peut-être l'une des causes profondes de cette lenteur du changement. La recherche sur les enquêtes doit faire appel à de nombreuses théories et méthodologies, dont la statistique, la technologie de l'information, la gestion, la communication et les sciences du comportement. Ces dernières sont nécessaires pour déterminer les causes fondamentales des erreurs non dues à l'échantillonnage. Si l'on se contente de quantifier ces erreurs, aucune amélioration n'est possible. À l'heure actuelle, les programmes de formation insistent sur les erreurs d'échantillonnage, de non-réponse et de couverture, et sur l'estimation en présence de ces erreurs. D'autres processus et sources d'erreur, tels que la production de mesures et le traitement des données, ne se voient pas accorder autant d'importance. D'où une situation où les études sur l'erreur de mesure et sur l'erreur de traitement des données sont rares comparativement à celles sur disons, la non-réponse. Tant dans le camp des producteurs que dans celui des utilisateurs, la confusion est importante en ce qui concerne les concepts et les méthodes. Une autre cause de lenteur pourrait être la règle du consensus appliquée dans certains organismes lorsqu'il s'agit de prendre des décisions concernant les changements. Cette règle repose sur le compromis. L'avis de nombreuses parties prenantes est recueilli et une décision est habituellement prise en se basant sur le plus petit commun dénominateur, ce qui n'est jamais une bonne norme. En outre, arriver à ce compromis de-mande habituellement beaucoup de temps et de ressources. Cette approche est très éloignée du modèle planifier-faire-contrôler-agir.

La qualité des enquêtes n'est pas une entité absolue. Le mode uniformisé de communication de l'information sur la qualité en vigueur à l'heure actuelle ne convient pas, puisque chaque utilisateur définit l'adéquation à l'usage prévu. Les décisions concernant des dimensions de la

Le processus d'évaluation a débuté par une auto-évaluation de chacun des huit produits clés. Les rapports de ces autoévaluations et d'autres documents pertinents ont été étudiés par des examinateurs externes qui ont ensuite rencontré les responsables des produits et leurs employés pour discuter des processus de production. Ensuite, les examinateurs ont présenté des évaluations détaillées et ont attribué une note à chaque produit. La procédure a permis de cerner d'importants domaines à améliorer, non seulement pour chaque produit, mais aussi pour l'ensemble des produits. Ce premier cycle d'évaluation a indiqué que l'erreur de mesure posait problème pour presque tous les produits clés. Comme toute autre approche de mesure ou d'indication de l'erreur d'enquête totale, celle-ci ne reflète pas vraiment l'erreur quadratique moyenne totale. Elle nécessite une description approfondie des processus et des améliorations apportées, et elle dépend fortement des compétences et des connaissances des examinateurs externes. Cette étude est présentée dans Biemer, Trewin, Japeac, Bergdahl et Pettersson (2012).

5.4 Profils de qualité

Dans le cas des enquêtes continues, il est possible d'élaborer des profils de qualité. Ce genre de document contient tout ce que l'on sait de la qualité d'une enquête continue ou d'un autre produit statistique assemblé au cours de plusieurs années. Les profils de qualité n'existent que pour quelques grandes enquêtes, qui sont toutes saut une, réalisées aux États-Unis, à savoir la Current Population Survey (Brooks et Bailar 1978), la Survey of Income and Program Participation (Jabine, King et Petroni 1990; Kalton, Winglee et Jabine 1998), la Schools and Staffing Survey (Kalton, Winglee, Kravchuk et Levine 2000), et l'American Housing Survey (Chakrabarty et Torres 1996). Fait exception la British Household Panel Survey (Lynn 2003). Le principal problème que pose un profil de qualité est son manque d'actualité, puisqu'il s'agit d'une complation des résultats d'études de la qualité qui prennent souvent beaucoup de temps. L'objectif du profil de qualité est de cerner les domaines où ils existent des lacunes dans les connaissances sur les erreurs, afin de pouvoir apporter des améliorations. Kasprzyk et Kalton (2001) ainsi que Doyle et Clark (2001) passent en revue l'utilisation des profils de qualité aux États-Unis.

6. Et maintenant, où allons-nous ?






Les notions de gestion de la qualité ont été influencées dans de nombreux organismes statistiques. Des concepts tels que le leadership, la culture de la qualité, la prévention des problèmes, la clientèle, la concurrence, l'évaluation des risques, la réflexion au sujet du processus, l'amélioration,

l'excellence opérationnelle et l'architecture opérationnelle sont des sujets abordés de plus en plus fréquemment par les dirigeants des organismes d'enquête, par exemple, Trewin (2001), Pink (2010), Fellegi (1996), Brackstone (1999), de Vries (1999), Groves (2011), et Bohata (2011). Le monde des enquêtes semble s'engager dans une direction où la production de statistiques devient rationnelle et rentable, mais l'évolution est lente. Certains organismes ont commencé à se servir d'un modèle de gestion de la qualité à des fins d'autoévaluation et d'orientation. Le modèle d'excellence de l'EFQM est celui qui est recommandé aux instituts nationaux de statistique qui font partie du Système statistique européen et de ceux d'entre eux, ceux de la République pour l'obtention du Prix national de l'EFQM de leur pays. Certaines entreprises de marketing sont certifiées d'après la norme ISO 9001 de management de la qualité et d'autres sont certifiées d'après la norme ISO 20252 concernant les études de marché et les études sociales et d'opinion. Ce développement devrait aboutir à des améliorations de la qualité, mais nous ne pourrions pas vraiment en être certains tant que nous ne commencerons pas à recueillir des données pertinentes. Cependant, une chose est sûre. Certains clients préfèrent les fournisseurs de services qui sont certifiés, qui ont gagné des prix ou qui peuvent donner la preuve que leur travail est conforme à un cadre ou à un modèle de qualité. Rares sont les clients qui jugeraient qu'un tel attribut est négatif.

Les marges d'erreur que nous associons aux estimations sont habituellement trop étroites, puisqu'elles n'englobent pas toutes les sources de variation. Les estimations ponctuelles peuvent se situer hors des limites à cause des biais. Idéalement, il serait utile de pouvoir produire des estimations de l'erreur d'enquête totale plutôt que celles produites aujourd'hui. Toutefois, ce genre de progrès n'est pas réaliste. Nous ne sommes pas en mesure de produire ce genre d'estimations, même à l'occasion, pour des raisons de budget, de temps et de méthodologie. Cela nous laisse les indicateurs de l'erreur d'enquête totale et de ses composantes. Ces indicateurs n'ont qu'une valeur limitée pour les utilisateurs, qui ne savent que faire des taux de réponse, de la variance de réponse mesurée par répétition des interviews ou des taux de rejets au contrôle. Par contre, ils sont très utiles pour les producteurs des données d'enquête. Par exemple, des études par répétition des interviews permettent de déceler la fausseté et les questions d'enquête pour lesquelles la réponse manque de cohérence. La majorité des utilisateurs apprécient la crédibilité du fournisseur de services et cette crédibilité tient en partie à la capacité de présenter des données exactes. Un autre aspect important de la crédibilité est la volonté qu'ont les fournisseurs d'évaluer leur propre qualité et de communiquer les résultats de ces

ministère des Finances de la Suède souhaite que les résultats des évaluations de la qualité permettent de suivre les améliorations de la qualité au fil du temps. Comme il faut évaluer la qualité d'un grand nombre d'enquêtes, de registres administratifs et d'autres programmes de l'organisme, il est nécessaire de disposer de certains indicateurs qui peuvent servir de mesures indirectes remplaçant les mesures réelles de la qualité. Parallèlement, le processus d'évaluation doit être complet, la communication des résultats doit être simple et les résultats doivent être crédibles. Pour chacune des sources d'erreur – spécification, base de sondage, non-réponse, mesure, traitement des données, échantillonnage, modélisation/estimation et révision –, huit produits clés ont été notés chacun comme étant mauvais, passable, bon, très bon ou excellent en ce qui concerne cinq critères. Ceux-ci étaient la connaissance des risques, la communication avec les utilisateurs, le respect des normes et des pratiques exemplaires, l'expertise disponible, et les réalisations en regard des plan d'amélioration des risques et/ou d'amélioration. Les lignes directrices de notation variaient selon le critère. Voici celles appliquées pour la connaissance des risques :

Exemple de lignes directrices de notation – Connaissance des

Mauvais		La documentation interne sur le programme ne mentionne pas la source d'erreur comme un facteur de risque possible pour la qualité du produit.	Mais : Aucun effort a été fait pour évaluer ces risques.	Mais : Les évaluations n'ont pas exploré les conséquences des erreurs sur les données. L'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.
Passable		La documentation interne sur le programme ne mentionne pas la source d'erreur comme un facteur de risque possible pour la qualité des données.	Mais : Les évaluations n'ont pas exploré les conséquences des erreurs sur les données. L'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.	Mais : Les évaluations n'ont pas exploré les conséquences des erreurs sur les données. L'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.
Bon		Un certain effort a été fait pour évaluer l'effet possible de la source d'erreur sur la qualité des données.	Mais : Les évaluations n'ont pas exploré les conséquences des erreurs sur les données. L'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.	Mais : Les évaluations n'ont pas exploré les conséquences des erreurs sur les données. L'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.
Très bon		Des études ont été menées pour estimer les pertes de la composantes du biais et de la variance associées à la source d'erreur et elles sont bien documentées.	Mais : Les évaluations n'ont pas exploré les conséquences des erreurs sur les données. L'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.	Mais : Les évaluations n'ont pas exploré les conséquences des erreurs sur les données. L'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.
Excellent		Il existe un programme permanent de recherche en vue d'évaluer toutes les composantes des TPE/QM associées de la source d'erreur, et leur incidence sur l'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.	Mais : Les évaluations n'ont pas exploré les conséquences des erreurs sur les données. L'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.	Mais : Les évaluations n'ont pas exploré les conséquences des erreurs sur les données. L'analyse des données. Le programme est bien conçu et axé sur les éléments approuvés, et l'information fournie requise pour faire face aux risques dus à cette source d'erreur.

diminué ou est-elle restée la même ? ». Lorsque les résultats ont été compilés pour ces trois catégories pour l'ensemble de l'organisme, il s'est avéré qu'une très faible proportion de gestionnaires avait déclaré une baisse de qualité, une proportion un peu plus élevée, une amélioration de la qualité, tandis qu'une vaste proportion avait déclaré qu'il n'y avait pas eu de changement. Les gestionnaires n'avaient tout simplement pas les moyens appropriés d'évaluer la qualité globale. En outre, des quantificateurs vagues, tels que « régulièrement », « utile » et « respect des normes », sont une invitation à fournir des évaluations gâtées. De surcroît, la plupart des gestionnaires ne veulent pas faire d'évaluation. Il est possible d'accroître la valeur de ces évaluations en posant des questions supplémentaires pour obtenir des renseignements détaillés sur la façon dont les travaux liés à la qualité ont été menés et à quel moment. Certains organismes recourent à des équipes internes qui vérifient les produits importants. Julien et Royce (2007) décrivent une vérification de la qualité de neuf produits menée par Statistique Canada afin de repérer les points faibles et leurs causes fondamentales, ainsi que pour dégager les pratiques exemplaires. Des équipes d'examen constituées de gestionnaires adjoints ont été créées afin que chaque examinateur passe en revue trois programmes différents. Le principal point faible d'une telle approche est l'élément interne proprement dit. Chaque examinateur sait que son tour d'être soumis à un examen viendra tôt ou tard, ce qui risque de le freiner. Le problème est également interne en ce sens que les utilisateurs ne sont pas explicitement présents durant le processus d'examen. Toutefois, dans son programme général de vérification de la gestion de la qualité des données, Statistique Canada insiste beaucoup sur son système de liaison avec les utilisateurs (Julien et Born 2006), qui est l'un des cinq systèmes formant le cadre d'assurance de la qualité de l'organisme, les autres étant la planification intégrée, les méthodes et les normes, la diffusion et la production de rapports sur les programmes.

Une autre variante de l'autoévaluation est celle où elle précède une vérification externe. Statistiques Netherlands (1997) décrit comment le Service des méthodes statistiques a été évalué par son personnel. L'évaluation a produit une liste de points faibles et de points forts qui a ensuite été examinée par une équipe externe. Habituellement, une vérification externe s'appuie pour l'évaluation sur certaines références, telles qu'un ensemble de règles, une norme ou un code de bonnes pratiques. La vérification aboutit alors à un certain nombre de recommandations concernant l'organisme ou le produit ou service en question. Récemment, Statistiques Sweden a élaboré un système général d'évaluation de l'erreur d'enquête totale. Le

trois phases. Durant la phase expérimentale, quelques options de plan de collecte sont mises à l'essai (par exemple, en ce qui concerne le niveau des mesures incitatives). Durant la phase principale de collecte de données, l'option choisie durant la phase expérimentale est mise en œuvre et se poursuit jusqu'à ce que soit atteinte la limite de capacité. Durant la phase de suivi des cas de non-réponse, des méthodes spéciales sont mises en œuvre pour réduire le biais de non-réponse et pour contrôler les coûts de la collecte des données. Ces méthodes comprennent le scénario d'échantillonnage double de Hansen-Hurwitz, l'augmentation des mesures incitatives et l'utilisation d'intervieweurs plus expérimentés. De nouveau, les efforts se poursuivent jusqu'à ce qu'une réduction supplémentaire du biais de non-réponse ne soit plus rentable. Le modèle des Six Sigma est le modèle d'excellence opérationnelle le plus élaboré, puisqu'il s'appuie fortement sur des méthodes statistiques. Il contient un jeu important de techniques et d'outils qui peuvent être utilisés pour contrôler et améliorer les processus. La conception et la mise en œuvre entièrement adaptées combinent les caractéristiques de contrôle de l'amélioration continue de la qualité, de la collecte adaptative et du modèle Six Sigma, afin de surveiller simultanément les multiples sources d'erreur. Biemer et Lyberg (2012) donnent plusieurs exemples de caractéristiques essentielles à la qualité et de mesures pour divers processus d'enquête. Par exemple, dans le cas du processus de mesure, les attributs qui sont des caractéristiques essentielles à la qualité pouraient inclure les aptitudes à repérer et à corriger les questions d'enquête qui posent problème, à déceler et à contrôler les erreurs de réponse, et à minimiser les biais et les variances dus aux intervieweurs. Les mesures correspondantes pourraient inclure le nombre d'éléments de données manquants par question, les taux de refus selon la taille de l'entreprise, les résultats des mesures répétées, les contrôles effectivement modifiés, et les résultats du travail sur le terrain par intervieweur. Les mesures peuvent être analysées en utilisant les méthodes de contrôle statistique des processus ou d'analyse de la variance. Différentes mesures connexes peuvent être affichées simultanément sous forme d'un tableau de bord. Par exemple, si l'une des caractéristiques essentielles à la qualité est la capacité de découvrir les intervieweurs qui trichent, nous pourrions créer un tableau de bord montrant la durée moyenne d'interview par intervieweur et la distribution de certaines caractéristiques de l'échantillon de nature sensible, également par intervieweur.

5.3 Autoévaluations et vérifications

Les principes de gestion de la qualité ont mené à l'introduction des concepts d'autoévaluation et de vérification dans la production de statistiques. Nous souhitions virement savoir ce que les utilisateurs, les clients, les propriétaires et d'autres parties prenantes pensent des produits et services fournis par l'organisme statistique. Un certain nombre d'outils sont disponibles pour ce genre d'évaluation. Nous avons déjà mentionné le sondage sur la satisfaction des clients. Les autres outils comprennent les sondages auprès des employés, les vérifications internes et les vérifications externes. Les sondages auprès des clients peuvent jeter de la lumière sur ce que les utilisateurs pensent des produits et services qui leur sont fournis. Ils peuvent servir à déterminer les besoins des utilisateurs et à cerner les caractéristiques du produit qui importent vraiment. Une autre série de questions pourrait avoir trait à l'image de l'organisme et à sa comparaison à celle d'autres organismes, qu'il s'agisse ou non de concurrents. Les sondages sur la satisfaction des clients sont très fréquents dans notre société. Souvent, il est impossible de les utiliser pour faire des inférences au sujet de la population cible d'utilisateurs, à cause de défauts méthodologiques et conceptuels. L'abondance de sondages sur la satisfaction, développés et mis en œuvre par des personnes ne possédant aucune formation officielle en méthodes d'enquête, contribue à l'accueil tiède qui leur est réservé dans des situations plus sévères dominant lieu à des erreurs de non-réponse et de mesure. Ainsi, le sondage sur la satisfaction des utilisateurs mené en 2007 par Eurostat comprenait deux sondages distincts. L'un, lancé sur la page Web d'Eurostat, avait pour population cible 3 800 utilisateurs inscrits. Seuls les utilisateurs inscrits qui sont entrés dans le site Web durant la période de collecte des données ont été exposés à la demande de participation au sondage, ce qui a donné un taux de réponse d'environ 5 %. Le second sondage, réalisé par courrier électronique, a été envoyé à plusieurs utilisateurs importants identifiés par Eurostat. Cet environnement plus contrôlé a produit un taux de réponse de 28 %. Ces sondages posent aussi des problèmes pour ce qui est d'identifier le répondant le plus approprié. Le choix du « mauvais » répondant au sein d'un organisme aboutira certainement à des résultats non éclairés et trompeurs.

Le type le plus simple d'autoévaluation est le questionnaire ou la liste de vérification remplie par le gestionnaire de l'enquête. Un exemple est offert par Statistics New Zealand. Il s'agit d'une liste de vérification comprenant un certain nombre d'indicateurs ou de déclarations, tels que « les besoins d'information sont évalués régulièrement en consultant les utilisateurs », « documentation utile et accessible », « production et surveillance régulières d'indicateurs de l'exactitude » et « respect des normes de présentation ». Le gestionnaire doit répondre par oui ou par non à chaque question et faire un commentaire s'il le juge nécessaire. Statistics Sweden utilisait un système similaire dont l'une des questions était « Comparativement à l'année dernière, la qualité globale de votre produit s'est-elle améliorée, a-t-elle

5.1 Estimations directes de l'erreur d'enquête totale

Les décompositions existantes de l'erreur quadratique

moyenne décrite, par exemple, dans Hansen et coll. (1964),

Fellegi (1964), Anderson, Kasper et Frankel (1979), Biemer

et Lyberg (2003), Weisberg (2005), et Groves et coll. (2009)

possibles de calculer directement l'EQM dans les situations

pratiques d'enquête, parce que ce calcul nécessite en général

une estimation des paramètres qui est essentiellement

exemple d'erreur. Toutefois, il est possible d'obtenir une

deuxième meilleure estimation de la vraie valeur des

paramètres si des ressources sont disponibles pour recueillir

des données par une méthode considérée comme la « norme

de référence », mais qui n'est ni de coût abordable ni pra-

tique dans des conditions normales d'enquête. Il s'agit de la

méthode classique d'évaluation lorsque la valeur vraie des

paramètres peut être définie de manière unique. Les

méthodes considérées comme la norme de référence sont

très rares, parce qu'elles sont exigeantes et que leur valeur

est parfois mise en doute (Nations Unies 2010). Par ailleurs,

des études d'évaluation régulières à plus petite échelle sont

nécessaires pour obtenir des indices quant aux problèmes

opérationnels et méthodologiques.

5.2 Indicateurs de qualité

La communication continue de l'erreur d'enquête totale est une tâche gigantesque qu'aucun organisme d'enquête n'entreprend. Ils fournissent plutôt des indicateurs ou des déclarations concernant la qualité. Par exemple, selon le manuel de production des rapports sur la qualité d'Eurostat (2009a), il convient de mesurer les indicateurs suivants :

- coefficient de variation ;
- taux de surcouverture ;
- taux de rejets à la vérification ;
- taux de réponses totales ;
- taux de réponses partielles ;

- taux d'imputation ;
- nombre d'erreurs ;
- ampleur moyenne des révisions.

Le thème commun ici est que ces éléments sommaires des paradosées sont des indicateurs qui peuvent être calculés sans effectuer d'études spéciales. Le jeu d'indicateurs qui peuvent être calculés directement d'après les données est douteux. Par exemple, inclure la surcouverture, mais non la sous-couverture, simplement parce que la première peut être calculée directement d'après les données disponibles, n'est pas logique. C'est la sous-couverture qui pose le problème de couverture le plus important dans les enquêtes. Le manuel prescrit certes que le producteur évalue les biais possibles (tant son signe que sa grandeur), mais il ne décrit pas clairement comment cela doit se faire. Il est demandé au producteur d'inclure les résultats des évaluations et des contrôles de la qualité, si cette information existe aussi. Les mesures du niveau d'effort pour des processus tels que la conception des questionnaires et la formation des coders seraient les bienvenues. Il n'existe aucun format normalisé de présentation de cette information qualitative et quantitative. De toute façon, la liste d'indicateurs clés est très restreinte si on la compare à la liste complète des principales sources d'erreur, et il est difficile de voir comment ces indicateurs sont perçus par les utilisateurs et comment ils peuvent être utilisés par le producteur pour améliorer le processus.

Le producteur a besoin d'une liste plus complète d'indicateurs pour pouvoir mesurer ou évaluer divers niveaux de qualité pour s'assurer que la mise en œuvre du plan d'enquête est maîtrisée ou être capable de mettre sur pied un projet d'amélioration de la qualité. Le plan d'enquête initial doit être modifié ou adapté durant la mise en œuvre afin de contrôler les coûts et de maximiser la qualité. Biemer (2010) discute de quatre stratégies de réduction des coûts et des erreurs en temps réel, à savoir l'amélioration continue de la qualité (ACQ), la collecte adaptative (Groves et Heeringa 2006), les Six Sigma (Bryfyogle 2003), ainsi que la conception et la mise en œuvre totalement adaptées.

Pour appliquer la stratégie d'amélioration continue de la qualité, il faut déterminer quelles sont les variables clés du processus ainsi que les caractéristiques de ce dernier qui sont essentielles à la qualité. Il faut élaborer des mesures fiables, essentielles à la qualité, du coût et de la qualité. Les mesures sont en temps réel, du coût et de la qualité. Les mesures sont surveillées en permanence durant le processus et des interventions ont lieu afin de s'assurer que les coûts et la qualité demeurent dans les limites acceptables. La stratégie de collecte adaptative a été conçue pour réduire le biais de non-réponse dans les entrevues en personne. Elle comprend

statistique des processus. Aucun autre organisme ne semble le faire.

L'édition la plus récente de ces lignes directrices est donnée dans Statistique Canada (2009).

De nombreux autres organismes statistiques de par le monde possèdent leurs propres normes de qualité. Elles sont parfois décrites comme des lignes directrices ou des normes et parfois, comme des systèmes de soutien opérationnel ou des cadres d'assurance de la qualité. Quoi qu'il en soit, le contenu et le style varient d'un organisme à l'autre, mais il faut que la variation soit gérable. Il devrait être possible d'arriver mondialement à un plus haut degré de normalisation, puisque cela s'est fait dans d'autres domaines, tels que les voyages aériens. Apfeld, Cartuthers, Lee, Oehm et Yu (2011) discutent de production de statistiques à l'Australian Bureau of Statistics.

La question est de savoir si des normes internationales amélioreraient la qualité des enquêtes en général. Certains domaines dans lesquels des normes seraient avantageuses comprennent le calcul des indicateurs de qualité utilisés fréquemment, tels que les taux d'erreur et les effets de plan, ainsi que les pratiques exemplaires pour la traduction du matériel d'enquête, le traitement des enquêtes ne parlant pas la langue du pays, et la pondération pour tenir compte de la non-réponse. Il ne faut pas oublier que, quand une norme est mise, elle doit être mise à jour continuellement et qu'il est bien connu qu'elles sont parfois difficiles à appliquer. Si les normes sont exhaustives, le praticien peut se sentir écrasé et par conséquent, les ignorer en grande partie, à moins que leur application ne soit rendue obligatoire et vérifiée.

4.6 Modèles de processus opérationnels statistiques

Au cours des dernières années, des concepts tels que les modèles de processus opérationnels et l'architecture opérationnelle ont été intégrés par certains organismes statistiques dans les travaux concernant la qualité. Afin de rendre les processus de production plus efficaces et plus souples, on peut les percevoir comme faisant partie d'un modèle d'architecture opérationnelle (Reedman et Julien 2010). Dans le domaine de la production statistique est élaboré générique du processus de production statistique est élaboré conjointement par la CEE-ONU, Eurostat et l'OCDE. Tout remaniement de système doit être dicté par les demandes des clients, les évaluations des risques et les nouveaux développements. Les principes architecturaux qui sous-tendent cette école de pensée sont résumés dans Doherty (2010), qui discute du renouvellement de l'architecture à Statistique Canada.

5. Mesure de la qualité

Donc, la qualité est un concept multidimensionnel et sa mesure est une tâche compliquée. Nous avons noté que la qualité des enquêtes peut être considérée comme un concept sous-jacents qui mènent au produit et à l'organisation qui fournit les moyens d'exécuter les processus et de livrer le produit ou le service avec succès. Il existe essentiellement deux façons de mesurer la qualité. L'une est l'estimation directe de l'erreur d'enquête totale ou de certaines composantes de cette dernière. L'autre consiste à mesurer des indicateurs de la qualité dans l'espoir qu'ils reflètent efficacement le concept proprement dit.

Voici quelques-uns des principes :

- La prise de décisions doit être optimale à l'échelle de l'organisme, ce qui implique la centralisation de l'information, du soutien méthodologique et du traitement des données.
- L'utilisation de services intégrés, tels que la collecte, la saisie et la diffusion des données, doit être optimisée.
- La réutilisation doit être maximisée et prévoir le plus petit nombre possible de processus opérationnels distincts et le plus petit nombre possible de systèmes informatiques.
- La boîte à outils de l'organisme doit être réduite au minimum.
- Le personnel doit bien connaître les outils et les systèmes.
- Les reprises, telles que les vérifications répétées, doivent être éliminées.
- Les activités de base doivent être le point de concentration, et le travail lié au processus de soutien doit être externalisé.
- Les activités de développement doivent être séparées des opérations continues.
- La collecte électronique des données doit être considérée comme le mode initial.
- Les obstacles structureaux, tels que le chevauchement ou le manque de clarté des mandats, doivent être éliminés.

très subjective, comme « l'étendue, la précision et le coût des statistiques européennes sont proportionnés aux besoins », tandis que d'autres sont plus spécifiques, tels que « un horaire standard de diffusion des statistiques européennes est porté à la connaissance du public ». Des exemples par les pairs de la conformité à un ensemble de principes menés en utilisant une version antérieure du Code ont révélé, ce qui n'est pas surprenant, que de nombreux organismes statistiques nationaux en Europe ont de la difficulté à les respecter (Eurostat 2011a). Par conséquent, afin de faciliter la mise en œuvre du Code, on a élaboré un cadre de soutien, appelé cadre d'assurance de la qualité (CAQ) qui contient des directives plus spécifiques concernant les méthodes et les références (Eurostat 2011b). Ce cadre semble être un document fort utile, car ses références sont principalement des résumés de l'état des connaissances dans des domaines tels que l'échantillonnage, la conception de questionnaires, la vérification, et ainsi de suite qui encourage la conformité aux pratiques exemplaires courantes.

Le Code de bonnes pratiques présente de nombreuses similitudes avec les Principes fondamentaux de la statistique officielle de l'ONU (de Vries 1999). Ces principes professionnels, qui constituent, en grande partie, un élément fait défaut dans le développement actuel de la production de statistiques (Kotz 2005). Même des pays voisins peuvent suivre des approches très différentes et posséder des niveaux de compétence en méthodologie très différents, et les différences sont parfois difficiles à expliquer. Nous savons par expérience que la collaboration en matière de développement est difficile à réaliser. Nous nous rendissons compte qu'il est difficile d'appliquer des idées qui pourraient s'adapter à nos systèmes. En revanche, il est plus difficile de se mettre d'accord sur des approches communes. Une norme globale qui se rapporte à la production de statistiques est l'ISO 20252 – Etudes de marché, études sociales et d'opinion (Organisation internationale de normalisation 2006). Il s'agit d'une norme de procédure comportant environ 500 exigences concernant les activités de recherche au sein d'un organisme. Il s'agit d'une norme minimale portant sur ce qu'il faut faire plutôt que sur la façon dont il faut faire les choses. Elle est appropriée pour les organismes qui réalisent des enquêtes et ils peuvent faire une demande de certification, dont la plupart étaient des entreprises de marketing, ont été certifiées. Un organisme statistique national (Uruguay) a été certifié en 2009 et Statistics Sweden prévoit obtenir la certification en 2013, mais ces organismes nationaux sont les seuls à s'être engagés sur cette voie. La norme porte sur le système de gestion de la qualité de l'organisme, ainsi que la gestion des éléments excédents de la recherche, la collecte

des données, la gestion et le traitement des données, ainsi que la production de rapports sur les projets de recherche (Blyth 2012).

Les normes du système statistique fédéral des États-Unis (OMB 2002) dont l'objectif était d'assurer et de maximiser la qualité, l'objectivité, l'utilité et l'intégrité de l'information diffusée par les organismes fédéraux. L'OMB (2006a) a également émis des normes et des lignes directrices pour les enquêtes. Elles sont bâties d'une manière classique. Vient d'abord une norme telle que « les taux de réponse doivent être calculés en utilisant des formules normalisées pour mesurer la proportion de l'échantillon admissible qui est représentée par les unités répondantes dans chaque étude, à titre d'indicateur du biais de non-réponse possible. » Cette norme est suivie d'un certain nombre de lignes directrices indiquant comment faire les calculs nécessaires, tandis que la dernière de ces lignes directrices précise que « si le taux de non-réponse globale dépasse 20 %, une analyse du biais manquant entièrement au hasard. » Comme dans le cas des normes du SSE, les lignes directrices de l'OMB sont complétées par un document de soutien (OMB 2006b) pour faciliter le respect des normes.

Dans le système statistique fédéral décentralisé des États-Unis, la plupart des organismes ont produit des documents dans lesquels sont adaptées les lignes directrices de l'OMB. Par exemple, l'U.S. Census Bureau possède ses propres normes de qualité statistique dont le niveau de détails techniques est plus élevé que celui des documents de l'OMB. Chaque norme est décrite au moyen d'exigences et de sous-exigences, et le document fournit souvent des exemples très spécifiques d'études qui peuvent être réalisées. Le National Center for Health Statistics, le National Center for Education Statistics, et l'Energy Information Administration sont d'autres exemples d'organismes américains dotés de normes concernant la qualité de l'information diffusée. Toutes ces normes peuvent être téléchargées à partir des sites Web de ces organismes.

Statistique Canada a émis des lignes directrices concernant la qualité depuis 1985. Elles sont similaires à celles du SSE puisqu'elles ne se limitent pas à mettre l'accent sur l'exacrité. Toutefois, elles sont nettement plus détaillées et contiennent un grand nombre de références. Une caractéristique particulière est que, pour certains processus, les lignes directrices prescrivent l'utilisation du contrôle

- Une pratique recommandée est une pratique privilégiée, mais il n'est pas obligatoire d'y adhérer.

Les catégories de cette classification ne sont certes pas mutuellement exclusives, surtout si l'on tient compte également des aspects linguistiques et culturels. Par exemple, en suédois, les politiques et les lignes directrices sont conçues très proches. Wikipedia, encyclopédie libre mais fondée sur un consensus, dit que les politiques découlent des normes, tandis que les lignes directrices découlent des pratiques exemplaires qui permettent de suivre ces normes. Cette phrase contient trois des catégories mentionnées par Colledge et March. Le mieux est probablement de traiter ces divers types de documents de la même façon. Ils visent tous à améliorer la qualité en réduisant divers types de variations et nous ne devons pas nous attarder trop sur la façon dont ils sont appelés.

Bien que les normes aient occupé une place importante dans la méthodologie d'enquête depuis longtemps, leur rôle s'est accru depuis que les organismes statistiques ont commencé à s'intéresser à la gestion de la qualité. Les premières normes, comme celles de Hansen et coll. (1967) et du U.S. Bureau of the Census (1974), étaient axées sur la discussion de la présentation des erreurs dans les données.

Au U.S. Census Bureau, toutes les publications doivent informer les utilisateurs que les données sont sujettes à erreur, que l'analyse pourrait être affectée par ces erreurs et que les erreurs d'échantillonnage estimées sont plus faibles que les erreurs totales. Dans le cas des grandes enquêtes, les erreurs non dues à l'échantillonnage doivent être traitées de manière plus détaillée, contrairement à ce qui se faisait dans le passé. De nombreux organismes statistiques ont adopté cette façon de penser. Par exemple, les cadres de la qualité mentionnés précédemment sont des extensions qui englobent d'autres dimensions de la qualité que l'exactitude. Le Système statistique européen a élaboré et lancé successivement ce qu'on a d'abord appelé les *Model Quality Reports* (rapports modèles sur la qualité), qui sont devenus aujourd'hui simplement la *Standard for Quality Reports* (norme pour les rapports sur la qualité) (Eurostat 2009a). La norme formule à l'intention des instituts nationaux de statistiques européens (noter la complexité conceptuelle) des recommandations pour la préparation de rapports sur la qualité pour une gamme « complète » de processus statistiques et de leurs produits. La norme traite des dimensions fondamentales de la qualité, à savoir la pertinence, l'exactitude, l'actualité, l'accessibilité, la cohérence et la comparabilité.

Examinons certains exemples. En ce qui concerne l'erreur de mesure, qui fait partie de la composante d'exactitude, la norme dit qu'un rapport sur la qualité doit contenir l'information suivante :

- la détection et l'évaluation générale des principaux risques en ce qui trait à l'erreur de mesure ;
 - si elles sont disponibles, les évaluations fondées sur des comparaisons à des données externes, la répétition des interviews ou des expériences ;
 - l'information sur les taux de rejet durant la vérification des données ;
 - les efforts déployés pour concevoir et mettre à l'essai les questionnaires, l'information sur la formation des intervieweurs et d'autres travaux sur la réduction des erreurs ;
 - les questionnaires utilisés qui doivent être annexés au rapport sous une forme ou l'autre.
- En ce qui concerne l'actualité, la norme dit que les rapports doivent contenir l'information suivante :
- pour les diffusions de données annuelles ou moins fréquentes : la durée moyenne de production pour chaque diffusion de données ;
 - pour les diffusions de données annuelles ou plus fréquentes : le pourcentage de diffusions effectuées à temps, selon les dates de diffusion planifiées ;
 - les raisons des diffusions tardives.

La norme comprend aussi des sections sur la façon de communiquer l'information au sujet des compromis entre les diverses dimensions de la qualité, l'évaluation des besoins et les perceptions des utilisateurs, le rendement et le coût, le fardeau de réponse, ainsi que la confidentialité, la transparence et la sécurité. Bien qu'elles comprennent une section sur les besoins et les perceptions des utilisateurs, ces derniers n'ont manifestement pas participé à la préparation de la norme proprement dite. Nous n'en savons toujours que fort peu sur la façon dont les utilisateurs perçoivent et utilisent l'information au sujet de la qualité. La norme est appuyée par un manuel beaucoup plus détaillé concernant les rapports sur la qualité (Eurostat 2009b) et les deux documents s'articulent autour des 15 principes énumérés dans le Code de bonnes pratiques de la statistique européenne, qui constitue le cadre de qualité fondamental pour le Système statistique européen. Les principes du Code de bonnes pratiques ont trait à l'indépendance professionnelle, l'adéquation des données, l'engagement sur la qualité, le secret statistique, l'impartialité et l'objectivité, une méthodologie solide, des procédures statistiques adaptées, la limitation du fardeau imposé au répondant, le rapport coût-efficacité, la pertinence, l'exactitude et la fiabilité, l'actualité et la ponctualité, la cohérence et la comparabilité, et enfin, l'accessibilité et la clarté. Chaque principe est accompagné d'un ensemble d'indicateurs que les organismes individuels peuvent mesurer pour déterminer s'ils sont ou non en conformité avec le Code. Certains indicateurs sont vagues et de nature

possibles des méthodes et de la technologie, et permet de proposer des améliorations des processus.

Parfois, les méthodes de contrôle statistique des processus peuvent être appliquées, mais parfois d'autres techniques analytiques sont nécessaires. Par exemple, pour pouvoir contrôler la falsification des données par les intervieweurs, il se pourrait que l'on doive examiner plusieurs processus simultanément, mais que la théorie et la méthodologie pour appuyer cette analyse ne soit pas directement disponible.

L'usage étargi de microdonnées qui ont trait à des enregistrements individuels, telles que les données sur les touches frappées et les enregistrements marqués d'un indicateur d'imputation, découle de l'utilisation des nouvelles technologies. Les procédures de collecte de données modernes produisent d'énormes quantités de ces types de paradoxes, tout comme le font aussi les systèmes de codage entièrement automatisé, ainsi que les systèmes de balayage optique des données. Il n'est pas logique de limiter le concept à la collecte des données.

4.5 Normalisation et outils similaires

Un moyen de maintenir la qualité du processus sous contrôle consiste à réduire la variation en favorisant l'utilisation de normes et de documents similaires. Colledge et March (1997) discutent de quatre catégories de documents.

- Une norme est un document qui doit être respecté presque sans exception. Les écarts par rapport à la norme sont déconseillés et requièrent l'approbation de la haute direction. Des mesures correctives doivent être prises lorsqu'une norme n'est pas entièrement satisfaite. Un organisme peut obtenir une certification de conformité à une norme. Il en est ainsi des normes ISO dont quelques-unes sont pertinentes pour les organismes statistiques.
- Une politique doit être appliquée sans exception. Par exemple, un organisme peut avoir une politique concernant l'utilisation de mesures incitatives pour accroître les taux de réponse.
- Plusieurs organismes ont élaboré des lignes directrices pour différents aspects de la production de statistiques. Habituellement, les lignes directrices peuvent être outrepassées s'il y a de « bonnes » raisons de le faire.

Les paradoxes ont reçu un nom génial et elles sont nécessaires pour juger de la qualité du processus. Cependant, la prudence est de rigueur. On ne doit jamais recueillir dans le processus et il est important de savoir comment les analyser.

Ce chapitre du livre de Morganstein et Marker a nettement influencé les travaux portant sur la qualité et la réflexion concernant les processus dans de nombreux organismes statistiques européens. L'intérêt pour ces questions s'est accru et certains organismes ont lancé leur propre système de gestion de la qualité dont l'amélioration des processus était un élément central.

Aux Joint Statistical Meetings de 1998, Mick Couper a donné un exposé sollicité sur la mesure de la qualité dans un environnement de collecte de données d'enquête assistée par ordinateur (CASIC). Il a mentionné que la nouvelle technologie produisait une foule de données secondaires susceptibles d'être utilisées pour améliorer le processus de collecte des données. Il a donné à ces données secondaires le nom de paradoxes, non pas dans son article, mais dans son exposé. Ce nom a été adopté très rapidement dans le monde des enquêtes et il était logique de définir la triologie des données, des métadonnées et des paradoxes. Donc, nous disposons d'un terme pour les données au sujet des données (métadonnées) et un autre pour les données au sujet du processus (paradoxes). Les paradoxes sont maintes-foisons des données sur le processus, mais très longtemps elles ont été limitées aux données au sujet du processus de collecte des données, alors que le terme utilisé dans de nombreux organismes statistiques européens était « données sur le processus » et tenaient compte de tous les processus sur le processus (Aitken, Hömgren, Jones, Lewis et Zilhao 2004). Récemment, on a assisté à un nouvel élargissement de la signification du concept. Kenickell, Mulrow et Scheuren (2009) nous rappellent ce qu'ils nomment macro-paradoxes, c'est-à-dire les données sur le processus global, tels que les taux de réponse, les taux de couverture, les taux de rejet et les taux d'erreur de codage, qui ont toujours été des indicateurs de la qualité du processus dans les organismes statistiques. Lyberg et Couper (2005), Kreuter, Couper et Lyberg (2010), et Smith (2011) emploient aussi la signification plus inclusive des paradoxes, qui tient compte d'autres processus que la collecte des données. Il existe un risque que, comme celui de qualité, le concept de paradoxes soit utilisé exagérément. On trouve des exemples de discussions dans lesquelles toutes les données, sauf les estimations d'après les données d'enquête, sont considérées comme des paradoxes, ce qui, naturellement, n'a aucun sens.

organismes statistiques et constaté que, dans la plupart des cas, la réflexion au sujet de la qualité n'était pas très avancée. Ils ont fondé leur plan générique sur leurs expériences pratiques et sur les notions générales de gestion de la qualité, exposées entre autres par Juran (1988), Deming (1986), Box (1990), et Scholes, Joiner et Strubel (1996). Ce plan compte essentiellement sept étapes :

- les caractéristiques critiques du produit sont précisées en collaboration avec l'utilisateur, en ce qui concerne les besoins généraux ainsi que les besoins plus uniques ;
- un schéma du déroulement du processus est élaboré par une équipe bien au courant. Le schéma doit comprendre la séquence des étapes du processus, les points de décision et les clients pour chaque étape ;
- les variables clés du processus sont identifiées parmi un ensemble plus grand de variables du processus ;
- la capacité de mesure est évaluée. Il est important que les décisions soient fondées sur de bonnes données et non pas simplement sur des données. Celles qui sont disponibles pourraient être inutiles. Il s'agit d'un domaine où les organismes statistiques sont avantagés par rapport à d'autres organismes. On ne doit pas tirer de conclusion au sujet de la stabilité du processus sans disposer d'information sur les erreurs de mesure. Avant tout, les données doivent permettre de quantifier l'amélioration ;
- la stabilité du processus est déterminée. Le schéma de variabilité des données sur le processus est analysé en utilisant des cartes de contrôle et d'autres outils statistiques ;

- la capacité du système est déterminée. Si la stabilité n'est pas atteinte après que la variation due à des causes spéciales a été éliminée, un effort d'amélioration est nécessaire. Des modifications peuvent être apportées au système lorsque la variation du processus est tellement grande que les spécifications, telles que les taux d'erreur minimaux ou les échéances de production, ne sont pas satisfaites. Des méthodes typiques en vue de réduire la variation sont l'élaboration et la mise en œuvre d'un nouveau programme de formation ou la mise en application d'une procédure opérationnelle normalisée. Cette dernière peut être une norme de processus, une norme relative aux meilleures méthodes courantes ou une simple liste de vérification ;
- la dernière étape du plan d'amélioration consiste à établir un système de surveillance permanente du processus. Nous ne pouvons pas nous attendre à ce que les processus restent stables au fil du temps. Pour de nombreuses raisons, une dérive s'amorce habituellement après un certain temps. Un système de surveillance facilite le suivi des nouvelles structures d'erreur, des nouvelles exigences des clients et des améliorations

production des statistiques. Considérer le processus de production comme une série d'actions ou d'étapes en vue d'atteindre un but particulier qui satisfait l'utilisateur mène à un produit de bonne qualité. La qualité du processus est évaluée en déterminant dans quelle mesure chaque étape satisfait à des exigences ou à des spécifications définies. Un moyen de contrôler la qualité du processus consiste à recueillir des données sur le processus qui peuvent varier avec chaque répétition de celui-ci. Les variables du processus qu'il est intéressant de surveiller sont celles qui ont un effet important sur le résultat final du processus. Donc, afin de vérifier la stabilité et la variation d'un processus, nous avons besoin de mécanismes pour cerner les variables clés et pour recueillir et analyser des données sur ces variables. La science de la gestion de la qualité nous a donné des outils tels que le diagramme en arêtes de poisson d'Ishikawa pour déterminer quelles pourraient être les variables clés du processus. La méthodologie de contrôle statistique des processus nous a donné des outils pour faire la distinction entre la variation dont la cause est spéciale et celle dont la cause est ordinaire et déterminer comment traiter ces deux types de variation. Habituellement, nous nous servons de cartes de contrôle qui ont été développées au départ par Shewhart (Deming 1986 ; Mundyk, Burgess et Xiao 1996) pour faire ces distinctions. Ensuite, nous recourons de nouveau à des méthodes issues de la science de la gestion de la qualité pour ajuster le processus, au besoin. Les organigrammes, ou diagrammes de flux, les diagrammes de Pareto et d'autres moyens simples permettent à l'équipe de production de repérer les causes fondamentales des problèmes en sont des exemples Juran 1988).

Les données sur les processus ont été utilisées pour contrôler les processus employés dans la production de statistiques depuis les années 1940, d'abord au U.S. Census Bureau, puis à Statistique Canada et, dans une certaine mesure, dans d'autres organismes également. Les processus habituellement vérifiés comprennent le codage, la saisie et l'impression des données et les données sur les processus étaient principalement des taux d'erreur. Certains contrôles des processus utilisés par le U.S. Census Bureau étaient tellement compliqués et coûteux que leur valeur a été mise en doute (Lyberg 1981), surtout parce que les boucles de rétroaction qui y étaient associées étaient inefficaces et ne visaient pas toujours à déterminer les causes fondamentales des erreurs. Il était courant de blâmer les opérateurs pour les problèmes causés par les systèmes et aucun accent n'était mis sur l'amélioration continue de la qualité. À l'époque, la réflexion était davantage axée sur la vérification et la correction.

Morganstein et Marker (1997) ont conçu un plan générique d'amélioration continue du processus qui peut être appliqué à la production de statistiques. Depuis les années 1980, ils avaient travaillé dans de nombreux

possible de prévoir leurs besoins d'information et d'analyse. Souvent, on peut isoler un ou quelques utilisateurs principaux avec lesquels communiquer, mais bon nombre de problèmes ayant trait à la conception et à la qualité des enquêtes sont tellement compliqués qu'une grande majorité d'utilisateurs s'attendent à ce que le fournisseur de services leur livre un produit contenant la plus petite erreur possible. Hansen et Voight ont déclaré que l'exactitude devrait être d'un niveau suffisant pour éviter les problèmes d'interprétation. Aujourd'hui, il semble exister un consensus voulant que les utilisateurs recherchent des produits et des services dans lesquels ils peuvent avoir confiance, ce qui signifie que le fournisseur de services doit être crédible. La plupart des utilisateurs n'ont pas la possibilité de vérifier les niveaux d'exactitude. Les aspects dont un utilisateur moyen peut discuter sont les questions ayant trait, par exemple, à l'exactitude, à l'accessibilité et à la pertinence. Des discussions détaillées au sujet de questions techniques et de problèmes de compromis de conception entre l'exactitude et la comparabilité sont plus difficiles à obtenir.

Au cours des dernières décennies, l'utilisateur a effectivement pris plus d'importance. Certains organismes élaboreront avec un utilisateur ou un client important des ententes de niveau de service qui énumèreraient les exigences concernant le produit ou service final afin qu'une vérification puisse avoir lieu au moment de la livraison. De nombreux organismes qui réalisent des enquêtes auprès des entreprises ont créé des unités qui communiquent continuellement avec les entreprises les plus grandes, puisque leur participation et leur fourniture de données exactes sont absolument essentielles au processus d'estimation (Williamack, Nichols et Sudman 2002). Les grandes entreprises ne sont pas des utilisateurs au sens strict. Il s'agit de fournisseurs importants ayant souvent un intérêt dans les résultats de l'enquête. Un autre outil de communication est le sondage sur la satisfaction des clients. La valeur de ce genre de sondage est limitée en raison du phénomène d'acquiescement et de la difficulté à trouver un répondant bien informé qui est prêt à répondre. En outre, de nombreux sondages sur la satisfaction des clients s'appuient sur l'autosélection, de sorte qu'ils n'ont aucune valeur inférieure. Les résultats de ces sondages peuvent être vus seulement comme des listes de problèmes et de préoccupations dont font part certains clients. Cette information peut évidemment être fort utile, mais elle ne convient pas pour l'estimation. De nombreux organismes d'enquête réalisent maintenant des sondages auprès des utilisateurs en continu (Ecochard, Hahn et Junker 2008).

4.4 La vue du processus

L'approche de la gestion de la qualité a de nouveau mis en relief qu'il importe d'avoir une vue du processus dans la

pendant un certain temps afin de traiter les risques associés à la mise en œuvre d'une nouvelle méthode ou d'un nouveau système. Selon Groves (2011), la culture de la production et les utilisateurs ont eu le dernier mot au sujet de tout changement, du moins jusqu'à présent. Simultanément, l'innovation est de plus en plus nécessaire dans de nombreux systèmes de production et il existe des exemples d'organisations cloisonnées auxquelles il ne reste plus beaucoup de temps (avant de devoir changer), parce que les ressources pour maintenir leurs systèmes font tout simplement défaut. Donc, même en cas de résistance au changement, le manque de ressources et la concurrence feront en sorte que les organismes statistiques deviennent davantage axés sur les processus et plus efficaces. Réduire le nombre de systèmes et d'applications, et privilégier une plus grande normalisation semblent être l'une des voies d'avenir.

4.3 Le client/l'utilisateur

L'apparition des concepts de gestion de la qualité dans les organismes statistiques a rendu plus visibles les destinataires des produits et services statistiques. Les entreprises commerciales parlent toujours de leur client, tandis que les organismes gouvernementaux ont eu tendance à les appeler des utilisateurs. Quoi qu'il en soit, la prise en compte du fait que quelqu'un est censé utiliser les produits finaux ne semble pas avoir été évidente pour certains fournisseurs. Il faut admettre que l'utilisateur a été un interlocuteur depuis que l'industrie des enquêtes a vu le jour. Aux États-Unis, les conférences à l'intention des utilisateurs étaient déjà assez fréquentes il y a 50 ans (Dalenius 1968 ; Hansen et Voight 1967). Ainsi, durant six mois de 1965 à 1966, le U.S. Census Bureau a organisé 23 conférences à l'intention des utilisateurs à travers le pays et a aussi organisé des groupes consultatifs. De nombreux pays ont privilégié les contacts de nature consultative avec les utilisateurs. Les conférences à l'intention des utilisateurs ont encore lieu, mais l'apport des utilisateurs est maintenant complété par d'autres moyens, tels que des discussions publiques et des forums sur Internet. Il est rare que les utilisateurs aient participé directement à la planification et à la conception des enquêtes. Même lors des discussions au sujet de la qualité des données, les producteurs ont agi en tant que représentants des utilisateurs. Les cadres de la qualité en sont un bon exemple. Les dimensions de la qualité ont été définies de la façon dont ceux-ci perçoivent l'information sur la qualité est, nous ne savons pas si l'information sur la qualité que nous fournissons leur est utile (Dalenius 1985b). En fait, une supposition éclairée est que, souvent, elle ne l'est pas. Dans le cas de beaucoup d'enquêtes, les utilisateurs sont nombreux et parfois inconnus, et il n'est pas

4. Exemples de projet d'amélioration de la qualité dans les organismes statistiques

La présente section donne des exemples de projets entrepris par les organismes statistiques en raison de l'intérêt général pour la qualité que manifeste la société.

4.1 L'erreur d'enquête totale

L'aspect peut être le plus important qu'il convient de souligner est que le domaine de la recherche et du développement portant sur la conception et la mise en œuvre des enquêtes, l'échantillonnage et les erreurs non dues à l'échantillonnage, ainsi que les effets des erreurs sur l'analyse des données, demeure florissant. L'obtention de données entachées de faibles erreurs est l'objectif principal des organismes de bonne réputation, comme en témoigne la publication régulière de manuels sur la collecte des données, l'échantillonnage, la non-réponse, la conception des questionnaires, les erreurs de mesure et les études comparatives. De nouveaux manuels traitant de sujets tels que les enquêtes auprès des entreprises, la traduction du matériel d'enquête et les paradosnées sont en cours de rédaction en vue de combler les lacunes dans ces domaines. Des revues, dont le *Journal of Official Statistics*, *Techniques d'enquête* et *Survey Practice*, sont entièrement consacrées à des sujets liés à la production de statistiques au sens large. De nombreuses autres revues, telles que le *Public Opinion Quarterly*, le *Journal of the American Statistical Association* et le *Journal of the Royal Statistical Society*, consacrent beaucoup d'espace aux méthodes d'enquête. La *Wiley series in Survey Methodology* et les conférences connexes (sur les enquêtes par panel, les méthodes d'enquête téléphonique (deux), les erreurs de mesure, la qualité des processus, les enquêtes-entreprises, la mise à l'essai et l'évaluation des questionnaires, la collecte de données d'enquête assistée par ordinateur, la non-réponse et les enquêtes comparatives) a eu beaucoup de succès et il en est de même des ateliers continus sur la non-réponse et l'erreur d'enquête totale. Donc, les idées quant aux sources d'erreur particulières et à leur traitement ne font pas défaut. Certains domaines, tels que les erreurs de spécification, les erreurs de traitement des données et l'effet des erreurs sur l'analyse des données, sont, certes, sous-étudiés, mais, dans l'ensemble, l'élargissement des connaissances sur les erreurs d'enquête suscite un véritable intérêt. Le défi tient à la communication de ces connaissances aux personnes qui travaillent dans les organismes statistiques et à l'élaboration de principes de conception qui peuvent être appliqués pour améliorer la production des statistiques. Une fracture évidente existe entre ce qui est connu grâce à la recherche et ce qui est connu et appliqué dans les organismes statistiquement les capacités semble nécessaire de renforcer continuellement les capacités

4.2 Risque et gestion du risque

L'un des éléments de la gestion de la qualité qui a fait son entrée dans l'univers des enquêtes est le risque et la gestion de ce dernier. Eltinge (2011) parle même du risque d'enquête total (*Total Survey Risk*) comme alternative au paradigme de l'erreur d'enquête totale. L'identification et la gestion des risques est un volet important de la vérification interne moderne (Moeller 2005) et est peut-être le seul élément important absent des cadres de gestion de la qualité, tels que celui de l'EFQM. Une source d'erreur peut être jugée comme posant un plus grand risque qu'une autre et doit, par conséquent, être traitée avec plus de soin et de ressources. Par exemple, ne pas posséder de système efficace de contrôle de la divulgation statistique est considéré comme une situation très risquée. Historiquement, la divulgation illégale de données est très rare, mais lorsqu'elle a lieu, elle risque de saper toutes les tentatives ultérieures de collecte de données. Certaines décisions concernant la conception des enquêtes peuvent être considérées comme risquées. Par exemple, si nous choisissons une méthode de collecte des données qui n'est pas adaptée au sujet de l'enquête, nous pourrions obtenir des estimations qui s'écartent tellement de la vérité que les résultats seront inutiles. L'étude de comportements de caractère délicat par interview sur place ou par téléphone au lieu d'un questionnaire à remplir soi-même pourrait en être un exemple. Il existe également des risques techniques qui doivent être décelés et évalués. Ainsi, l'U.S. National Agricultural Statistical Service (Gleason 2011) possède, comme de nombreux autres organismes, des plans de reprise après sinistre. Groves (2011) et Dillman (1996) discutent tous deux des visions différentes des risques qui pourraient émaner de la culture de production et de la culture de recherche au sein d'un organisme statistique. Le changement s'opère généralement lentement dans ces organismes, mais s'opère généralement lentement dans ces organismes, pour de bonnes raisons. Le changement pourrait aboutir à un échec, tel qu'une mise en œuvre infructueuse, des coûts importants et une réduction de la comparabilité des données. Donc, dans un certain sens, tant les producteurs que les utilisateurs des données ont tendance à hésiter à adopter les changements proposés par les chercheurs et par les auteurs avec laquelle les changements ont lieu. Il est courant de produire des mesures parallèles

La liste de principes ressemble à d'autres listes qui ont été établies par l'ONU et d'autres organisations.

3.2.4 Certaines conséquences particulières pour les organismes statistiques

La plupart des organismes statistiques adoptent les principes de gestion de la qualité à des degrés divers et avec plus ou moins de succès. Comme l'on fait remarquer Collège et March (1993), il est possible d'énumérer plusieurs obstacles à la mise en œuvre de ces principes. Un organisme gouvernemental peut avoir de la difficulté à motiver son personnel au moyen de primes monétaires, puisque la façon dont l'argent des contribuables peut être dépensé fait l'objet de restrictions. La diversité des utilisateurs et des produits complique le dialogue entre le fournisseur de services et l'utilisateur, et comme il est mentionné plus haut, ni les utilisateurs ni d'ailleurs les fournisseurs des données ne sont entièrement familiers avec tous les biais et autres problèmes de qualité présents dans la production de statistiques. L'effet des erreurs sur les utilisations des données peut varier et est souvent inconnu. La situation se complique encore davantage du fait que, contrairement à ce que connaissent la plupart des autres entreprises, les fournisseurs des organismes statistiques ne sont pas très enthousiastes. Les fournisseurs des autres entreprises sont payés, tandis que ceux des organismes statistiques, c'est-à-dire les répondants, qui reçoivent rarement un incitatif monétaire, doivent être motivés.

Par ailleurs, les organismes statistiques ont un énorme avantage quand il s'agit d'appliquer les principes de gestion de la qualité. Un organisme statistique sait comment recueillir et analyser les données qui orientent les efforts d'amélioration. L'une des pierres angulaires des concepts de gestion de la qualité est que les décisions doivent être fondées sur des données, et souvent, les entreprises qui ne connaissent pas de l'appui des statisticiens ne sont pas au courant des problèmes de qualité des données qui peuvent avoir des répercussions sur les décisions qu'elles prennent. Néanmoins, dans l'ensemble, un organisme statistique n'est pas différent de toute autre entreprise et il lui est fort possible d'appliquer les concepts de gestion de la qualité afin d'améliorer tous les aspects de son travail.

Certains organismes statistiques ont utilisé des modèles d'excellence opérationnelle pour l'évaluation. L'office technique de statistique a été déclaré lauréat du Prix national de la qualité de la Tchécoslovaquie pour 2009 dans la catégorie Secteur public en se basant sur le modèle d'excellence de l'EFQM. Il a obtenu 464 points. Le Leadership Group-Qualité d'Eurostat a recommandé que les organismes statistiques nationaux européens utilisent le modèle de l'EFQM pour leurs travaux sur la qualité, et les organismes de la Finlande et de la Suède comptent parmi ceux qui l'ont fait. Depuis que le Leadership Group a publié son rapport en 2001 (voir Lyberg, Bergdahl, Blanc, Booleman, Grfinewald, Haworth, Japac, Jones, Kömer, Kömer, Lindén, Lundholm, Madaleno, Rademacher, Signore, Zilhao, Tzongas et van Brakel 2001), d'autres cadres et normes ont été élaborés. Le Système statistique européen a lancé son code de bonnes pratiques, qui compte un certain nombre de principes associés à des indicateurs. Pour certains de ces principes, cependant, les indicateurs constituent plutôt des éclaircissements

vivement critiquée par Deming (1986) et d'autres, mais peut représenter le seul moyen de contrôle disponible dans des situations où le roulement du personnel est élevé et que l'on ne dispose pas de suffisamment de temps pour attendre que les processus soient stables.

Les paradoxes globales (Scheuren 2001) sont des taux de « erreurs » de différentes sortes. Les taux de non-réponse, les taux d'erreur de codage, les taux d'erreur de balayage optique et les taux d'erreur de liste en sont des exemples. Dans le cas de certaines opérations, les taux d'erreur sont calculés en recourant à la vérification, ce qui signifie que l'opération est répétée d'une certaine façon. Il en est ainsi de l'opération de codage. Pour d'autres opérations, le calcul peut être fondé sur un schéma de classification, comme pour le calcul des taux de non-réponse. Ces paradoxes globales nous renseignent sur le processus. Il s'agit de statistiques sur les processus, c'est-à-dire des sommaires de données. Un taux de non-réponse élevé signale des problèmes dans le processus de collecte des données et un taux élevé d'erreur de codage signale des problèmes dans le processus de codage. Partant de ces données sommaires, il est parfois possible de faire la distinction entre la variation dont la cause est ordinaire et celle dont la cause est spéciale, et de décider de la mesure à prendre.

Certains processus normalisés peuvent être contrôlés au moyen de simples listes de vérification. Ces dernières sont très efficaces parce qu'il est crucial que chaque étape du processus soit accomplie, et ce, dans le bon ordre (Morganstein et Marker 1997). La préparation au décollage effectuée par les pilotes d'avion en est un exemple. Peu importe le nombre de fois qu'ils ont décollé, sans liste de vérification, le jour viendra où ils oublieront un élément. Dans le domaine de la production de statistiques, l'échantillonnage est un processus de ce genre, même si les conséquences de l'oubli d'un élément sont moins graves. Il se pourrait fort bien qu'un organisme statistique possède un processus normalisé de sélection des échantillons et qu'une liste de vérification puisse être utilisée comme directive de travail et instrument de contrôle.

Une sorte de liste de vérification peut être utilisée dans les processus plus créatifs, tels que le processus de conception globale d'enquête. Il est impossible de normaliser ce processus, mais il est possible de dresser la liste d'un certain nombre d'étapes critiques qui doivent toujours être accomplies. La liste ne nous dit pas comment les accomplir. Elle sert juste à rappeler qu'une étape particulière ne doit pas être omise ni oubliée. Morganstein et Marker (1997) discutent de ce genre de liste de vérification et les appellent (ainsi que les listes de vérification plus simples) meilleures méthodes courantes (MMC). Ils décrivent le processus d'élaboration des MMC et la façon dont ces dernières peuvent être utilisées pour réduire la variation des processus dans les

organismes statistiques. Ainsi, un organisme pourrait disposer de sept méthodes et systèmes d'imputation différents dans sa boîte à outils. Le maintien de ces sept systèmes coûte cher. Il est peu probable qu'ils soient tous aussi efficaces les uns que les autres. S'ils le sont, il n'est gu'une date d'expiration est associée à chacune d'elle.

Dans un certain sens, les MMC sont naturellement des « pratiques exemplaires ». De nombreux organismes souhaitent mettre en œuvre et utiliser de telles pratiques. Morganstein et Marker offrent un processus pour élaborer ces pratiques exemplaires et les tenir à jour. Ce processus est utile pour un organisme s'il est possible de maintenir à un niveau minimum la variation de la conception des processus. Il devient alors facile de former le personnel et de modifier le processus quand il devient instable ou que de nouvelles méthodes sont mises au point. Par ailleurs, si les MMC et d'autres normes ne sont pas mises en application fermement au sein d'un organisme, leur usage ne sera pas répandu et l'investissement ne sera pas rentable.

3.2.3 Qualité organisationnelle

Les cadres sont responsables de la qualité au sens le plus large. C'est l'organisme qui assure le leadership, le perfectionnement du personnel, les outils permettant d'établir de bonnes relations avec la clientèle, les investissements et le financement. Le domaine de la gestion de la qualité nous a donné des modèles d'excellence opérationnelle qui peuvent nous aider à évaluer nos organismes statistiques de la même façon que d'autres entreprises le sont. Les deux principaux modèles d'excellence opérationnelle sont ceux du Baldrige National Quality Program et de l'European Foundation for Quality Management (EFQM).

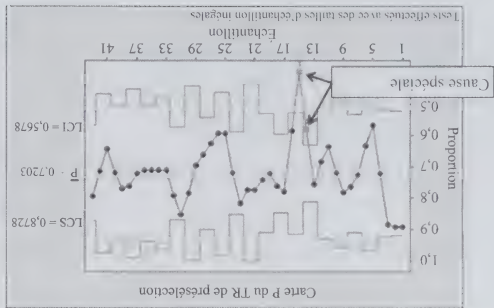
Ces modèles consistent en une liste de critères à vérifier pour évaluer un organisme. Les sept principaux critères utilisés pour décerner le prix Malcolm Baldrige sont le leadership, la planification stratégique, l'orientation client et le marché, l'information et l'analyse, l'orientation ressources humaines, la gestion des processus et les résultats opérationnels. Chaque critère comprend plusieurs sous-critères. Par exemple, l'orientation ressources humaines englobe les systèmes de travail, la formation et le perfectionnement des employés, et le bien-être et la satisfaction des employés. Les neuf critères sur lesquels s'appuie le modèle d'excellence de l'EFQM sont le leadership, les personnes, la stratégie, les partenariats et les ressources, les processus, produits et

3.2.2 Qualité des processus

Tous les processus doivent être conçus de façon qu'ils fournissent ce qu'ils sont supposés produire. Cela signifie qu'une certaine perspective d'assurance de la qualité est nécessaire lorsque les processus sont définis. Ainsi, le processus d'interview implique qu'un certain nombre d'éléments doivent être en place pour que le processus livre ce qui est attendu. Ces éléments sont, par exemple un bon choix d'intervieweur, ainsi que la mise en place d'un programme de formation, d'un système de rémunération, ainsi que les activités de supervision et de rétroaction. Donc, nous nous employons à intégrer la qualité dans le processus par la voie de l'assurance de la qualité. Les activités de contrôle de la qualité ne sont utilisées que pour vérifier si le processus fonctionne comme il est prévu. Elles ne peuvent, à elles seules, être utilisées pour intégrer la qualité dans le processus. Cette vision des processus est discutée plus en détail à la section 4.4. La qualité des processus est mesurée et contrôlée par sélection, observation et analyses des données sur le processus ou par sondages (Morganstein et Marker 1997 ; Couper 2005 ; Lyberg et Couper 2005). La théorie et les méthodes importées du domaine du contrôle statistique des processus peuvent aider le producteur à faire la distinction entre deux types de variation, à savoir la variation ordinaire et celle ayant une cause particulière. À condition que la variation totale soit contenue entre les limites supérieure et inférieure de contrôle associées aux cartes de contrôle choisies, le processus est déclaré être sous contrôle statistique et aucune amélioration n'est vraiment possible en essayant d'ajuster les résultats individuels. Des observations qui tombent en dehors des limites de contrôle (habituellement fixées à 3 sigma), sont une indication qu'il existe une cause spéciale de variation dont il faut s'occuper afin qu'après la correction, la variation soit ramenée à celle de cause ordinaire. La carte de contrôle P qui suit illustre une situation possible :

Un autre moyen de vérifier la qualité du processus consiste à utiliser l'échantillonnage d'acceptation (Schilling et Neubauer 2009), qui peut être appliqué à des situations où les éléments du processus peuvent être groupés par lots. Les lots sont contrôlés et, en fonction du résultat de ce contrôle, il est décidé si le lot doit être approuvé ou retravaillé. Les plans d'échantillonnage d'acceptation garantissent une qualité sortante moyenne en ce qui a trait à, disons, le taux de défaut, mais ne comportent aucune amélioration directe de la qualité. Il s'agit d'un instrument de contrôle qui convient pour des opérations telles que le codage, la vérification et le balayage optique, et ce, uniquement quand ces processus ne sont pas vraiment en contrôle statistique. La méthode a été

Donc, l'enchaînement des actions est le suivant. D'abord, on recherche l'origine des causes spéciales afin de pouvoir éliminer ce type de variation. Après cela, le processus ne présente plus que la variation ordinaire. Si cette variation est jugée trop grande, le processus doit être modifié. Les types de changements nécessaires sont parfois nécessaires pour réduire la variation du processus. Habituellement, il faut lancer un projet d'amélioration du processus et la littérature sur la gestion de la qualité propose un certain nombre d'outils utiles pour ce genre de projet. La plupart de ces outils sont empruntés à la statistique (cartes de contrôle, expériences, analyses de régression, diagramme de Pareto, nuages de points, stratification), mais il existe aussi des outils pour déterminer les causes probables du problème (diagrammes d'enchaînement du processus, séance de remue-méninges). Une opinion répandue est que les projets d'amélioration doivent être réalisés par les personnes qui travaillent avec le processus ou par des personnes très familières avec ce dernier par d'autres moyens. Parfois, nous parlons de créer une équipe d'amélioration, à laquelle participe aussi le client. Dans tout projet d'amélioration, les changements proposés doivent être testés. Quand Shewhart a développé ses premières cartes de contrôle, il a également soutenu que le travail d'amélioration doit suivre une série d'opérations qu'il a appelées *Plan-Do-Check-Act* (planifier, faire, contrôler, agir). Cette séquence nous indique que tout changement qu'il est proposé d'apporter à un processus doit être testé pour voir s'il améliore réellement le processus. Dans la négative, un autre changement est fait et les tests sont répétés. Deming a appelé ce courant de pensée le cycle de Shewhart, mais puisqu'il a passé beaucoup de temps à le promouvoir, nombreux sont ceux qui ont fini par l'appeler cycle de Deming. Les changements recherchés pourraient être une réduction de la variation du processus, une réduction des coûts, ou un accroissement de la satisfaction des clients. La méthodologie des projets d'amélioration est décrite, par exemple, dans Joiner (1994), Box et Fritsds (2006), Breyfogle (2003) et Deming (1986).



Carte de contrôle de processus avec valeurs extrêmes

(SSE) comme outil permettant aux instituts statistiques nationaux d'atteindre le niveau voulu de qualité organisationnelle. Le concept est qu'il n'est pas possible d'arriver à un produit de bonne qualité, selon les dimensions mentionnées (ou une autre définition de la qualité du produit), si l'organisme ne met pas en place de bons processus sous-jacents. On pourrait également soutenir que le moyen le plus efficace et le plus fiable d'obtenir un produit de bonne qualité est d'utiliser des processus de bonne qualité. Si nous considérons la qualité comme un concept à trois niveaux, elle peut être visualisée comme on le présente au tableau 2.

3.2.1 Qualité du produit

Les résultats qu'il est convenu de livrer sont appelés le produit. Il peut s'agir d'estimations, de jeux de données, d'analyses, de registres, de processus normalisés ou d'autre matériel d'enquête, tel que des bases de sondage et des questionnaires. La qualité du produit correspond au concept classique de qualité utilisé pour informer les utilisateurs ou les clients de la qualité du produit ou du service. Elle peut être mesurée et contrôlée par le degré de respect des spécifications et des exigences quant aux caractéristiques du produit qui forment les dimensions de qualité d'un cadre. Les mesures de l'exacitude et les marges d'erreur entrent dans cette catégorie. Sont également pertinentes les observations en vue de déterminer si les entités de niveau de service établies avec les clients ont été respectées. En harmonie avec les principes de gestion de la qualité, il est également assez fréquent de réaliser des sondages sur la satisfaction des utilisateurs afin de découvrir ce que ceux-ci pensent des produits et services fournis.

À la fin des années 1980 et au début des années 1990, de nombreux organismes statistiques se sont intéressés aux problèmes de qualité dépassant les aspects habituels de la qualité des données. Les questions concernant la satisfaction des clients, la communication avec les clients, la concurrence, la variabilité des processus, le coût de la main-d'œuvre, le gaspillage, les modèles d'excellence opérationnelle, les valeurs fondamentales, les pratiques exemplaires, l'assurance de la qualité, et l'amélioration continue de la qualité ont soudain fait partie des préoccupations quotidiennes de nombreux organismes.

Les organismes qui réussissent savent qu'il est nécessaire de s'améliorer continuellement (Kaizen) pour rester en activité et ils ont mis au point des mesures qui les aident à évoluer. Cela s'applique également aux producteurs de statistiques. Les changements qui sont censés améliorer le produit statistique sont déclenchés par les demandes des utilisateurs, par la concurrence des autres producteurs et par les valeurs des producteurs qui mettent l'accent sur l'amélioration continue en tant qu'environnement de fonctionnement général. Les mesures qui peuvent aider un organisme statistique à s'améliorer sont essentiellement les mêmes que pour les autres entreprises. Elles peuvent s'appuyer sur des modèles d'excellence opérationnelle, tels que celui de l'European Foundation for Quality Management (EFQM) (1999). Les valeurs fondamentales sur lesquelles repose le modèle de l'EFQM sont l'orientation résultats, l'orientation client, le leadership et la constance des objectifs, le management par les processus et les faits, le développement et l'implication des personnes, la formation continue, l'innovation et l'amélioration, le développement des partenariats, et la responsabilité sociale de l'organisme. Ce modèle a été adopté par le Système statistique européen

Tableau 2
Qualité - Concept à trois niveaux*

Niveau de qualité	Principaux intervenants	Instrument de contrôle	Mesures et indicateurs
Produit	Utilisateur, client	Spécifications du produit, ENS, études d'évaluation, cadres, normes	Cadres, conformité, EQM, sondage auprès des utilisateurs
Processus	Concepteur de l'enquête	CSP, cartes de contrôle, échantillonnage d'acceptation, analyse des risques, MMC, PON, paradosées, listes de vérification, vérification	Variation au moyen de cartes de contrôle, analyse d'autres paradosées, résultats des études d'évaluation périodique
Organisation	Organisation, propriétaire, société	Modèles d'excellence, ISO, CDP, examens, vérifications, autoévaluations	Scores, points forts et points faibles, sondages auprès des utilisateurs, sondages auprès du personnel

* ENS (Entente de niveau de service), CSP (contrôle statistiques des processus), MMC (meilleures méthodes courantes), PON (procédure opérationnelle normalisée) et CDP (code de bonnes pratiques du SSE).

En premier lieu, son élaboration n'a pas été précédée de communications avec les utilisateurs. Les producteurs de statistiques ont cru que les utilisateurs étaient intéressés par un ensemble particulier de dimensions, même s'il est évident que la vaste majorité d'entre eux pensent que les structures d'erreur sont trop difficiles à saisir et supposent que le producteur a la responsabilité de fournir les données les plus exactes possibles. Lorsque l'utilisateur ou le client a des exigences d'exactitude particulières, un dialogue plus approfondi peut s'établir entre eux. Selon les rares études importantes de la façon dont les utilisateurs perçoivent l'information sur la qualité, les utilisateurs s'intéressent surtout aux dimensions qui sont faciles à comprendre, telles que l'actualité et les indicateurs qui paraissent simples, comme les taux de réponse. L'utilisateur veut que l'organisme statistique producteur soit crédible, ce qui se traduit par la capacité de produire des données contenant des erreurs faibles ou du moins connues, et de les livrer en temps opportun, de manière fiable et accessible. L'idée qu'il serait possible de produire une mesure de la qualité totale fondée sur des évaluations pondérées des différentes dimensions n'est pas raisonnable, même si Mitrochic (1993) soutient le contraire. Dans son article, il présente des arguments en faveur d'un ensemble normalisé d'indicateurs de la qualité et donne un exemple hypothétique d'indicateurs de la qualité de données d'indexation et calculé un indice réel (dans cet exemple, les indicateurs sont la précision, la non-réponse, la fiabilité, l'actualité et les résidus). Même s'il était possible d'élaborer un indicateur composite sous forme d'un indice, l'utilisateur voudrait savoir quels indicateurs ont contribué le plus à la valeur de l'indice. Du point de vue de l'utilisateur, la valeur de l'indice la moins favorable pourrait encore refléter une situation offrant le plus haut niveau de qualité. Il est rare qu'une faible exactitude puisse être compensée par de bonnes évaluations sur d'autres dimensions, pas même dans le cas, lors des élections, des sondages faits à la sortie de l'isoloir, pour lesquels l'actualité est indispensable. L'exactitude demeure nécessaire et il est généralement reconnu que tous les organismes dignes de confiance doivent satisfaire aux normes d'exactitude (Scheuren 2001 ; Katon 2001 ; Brackstone 2001). Phipps et Fricke (2011) donnent un aperçu des cadres de qualité et de la littérature sur l'erreur d'enquête totale. Donc, nous pouvons convenir que la qualité des enquêtes est un concept multidimensionnel faisant intervenir plusieurs caractéristiques d'un produit ou service statistique.

3.2 Les répercussions du mouvement en faveur de la qualité sur les organismes statistiques

Simplement élargir le cadre de la qualité pour passer d'une ou deux dimensions à plusieurs d'entre elles ne suffit

Si l'exactitude est insuffisante, les autres dimensions sont à l'accessibilité (voir Weisman, Balyozov et Venter 2010), méthodologique, l'exactitude et la fiabilité, l'utilité, et dimensions de la qualité, à savoir l'intégrité, la rigueur comprend un ensemble de conditions préalables et cinq Statistics Sweden et de Statistique Canada. Le cadre du FMI ceux de l'OCDE, de l'Austrian Bureau of Statistics, de produits de qualité est un savant numéro d'équilibre dans lequel les utilisateurs informés jouent un rôle important. Des conflits existent habituellement entre l'actualité et l'exactitude, par exemple, à un suivi à grande échelle des cas de non-réponse. Un autre conflit est celui qui survient entre la comparabilité et l'exactitude, puisque l'application de nouvelles méthodes plus exactes pourrait perturber les comparaisons au fil du temps (Holt et Jones 1998).

Tableau 1
Cadre de qualité de l'OCDE

Dimension	Description
Pertinence	Les statistiques sont pertinentes si les besoins de l'utilisateur sont satisfaits.
Exactitude	Degré de rapprochement entre la valeur finalement retenue et la valeur réelle, mais inconnue, dans la population.
Credibilité	Le degré de confiance qu'ont les utilisateurs dans les produits de données en fonction de la perception qu'ils ont du fournisseur des données.
Actualité	Temps écoulé entre le moment où les données sont disponibles et le moment où elles deviennent.
Accessibilité	Facilité avec laquelle les données peuvent être localisées et consultées à l'intérieur des fonds de données.
Intelligibilité	Facilité avec laquelle l'utilisateur peut comprendre, utiliser et analyser correctement les données.
Cohérence	Reflète la mesure dans laquelle les produits de données sont reliés logiquement et mutuellement concordants.
Rentabilité	Une mesure des coûts et du fardeau imposé au fournisseur par rapport à la production.

Donc, de nombreux organismes ont adopté un concept de qualité multidimensionnel comprenant non seulement l'exactitude, mais aussi d'autres dimensions. Nous pourrions parler d'un vecteur de qualité dont les composantes varient légèrement d'un organisme à l'autre, tant en nombre qu'en contenu. Plusieurs problèmes sont associés à l'ap-proche du vecteur de qualité.

1900 à 1930 Les idées de Taylor sont appliquées, par exemple, aux chaînes de montage chez Ford et chez Mercedes Benz.

Années 1920 Fisher commence à élaborer des théories et des méthodes concernant les plans expérimentaux.

1924 Shewhart développe la carte de contrôle.

1940 Le U.S. War Department produit un guide pour l'analyse des données sur les processus.

1944 Deming présente la première classification des erreurs d'enquête.

1944 Dodge et Romig présentent la théorie et des tableaux pour l'échantillonnage d'acceptation.

1946 Deming part au Japon.

1950 Ishikawa propose le diagramme en arêtes de poisson comme outil pour déterminer les facteurs qui ont un profond effet sur le résultat du processus.

1954 Juran part au Japon.

1960 De nombreuses entreprises lancent un programme « zéro défaut ».

1960 Le U.S. Census Bureau élabore des programmes de contrôle de la qualité.

1961 Le U.S. Census Bureau lance son modèle d'enquête.

1965-1966 Kist et Slobodan Zarkovich commencent à parler de la qualité des données plutôt que des erreurs d'enquête.

Années 1970 De nombreux organismes statistiques fournissent des lignes directrices concernant la qualité totale (GQT).

1975 Lancement du cadre de la gestion de la qualité totale (GQT).

1976 Adoption par un organisme statistique du premier cadre de la qualité contenant plus de dimensions que la pertinence et l'exactitude.

1987 à 1989 Lancement de la norme ISO 9000, du prix Malcolm Baldrige, de la stratégie Six Sigma et des modèles de l'EFQM.

Années 1990 De nombreux organismes statistiques commencent à travailler avec des modèles d'amélioration de la qualité et d'excellence.

1997 Publication de la monographie sur la qualité des mesures et des processus d'enquête (*Survey Measurement and Process Quality*).

1998 Mick Couper introduit le concept des « données » en tant que sous-ensemble des données sur les processus.

À partir du milieu des années 1990, les principes de gestion de la qualité ont eu un immense effet sur de nombreux organismes statistiques. Il ne s'agit pas nécessairement d'un accroissement de la qualité à tous les niveaux (personne n'a vérifié ce fait). Mais les principes se sont traduits dans la plupart des organismes par une prise de conscience de l'importance du maintien de bons contacts avec les utilisateurs et avec les clients, et une aspiration à devenir « le meilleur » ou « de niveau international ». La qualité est à l'ordre du jour.

3.1 Le concept de qualité

Au cours des dernières décennies, il est devenu évident que l'exactitude et la pertinence sont des éléments nécessaires, mais non suffisants pour évaluer la qualité des enquêtes. D'autres dimensions sont également importantes pour les utilisateurs. L'élaboration de cadres de la qualité des enquêtes s'est déroulée principalement au sein des organismes de statistique officielle et a été déclenchée par le progrès technologique rapide et d'autres développements sociétaux. Les technologies de pointe ont créé des possibilités et suscité des demandes de la part des utilisateurs au sujet de dimensions éventuelles de la qualité, telles que l'accessibilité, l'actualité et la cohérence, qui n'étaient tout simplement pas mises en relief auparavant. Les décisions prises par la société sont devenues plus complexes et de portée plus mondiale, ce qui s'est traduit par des demandes de statistiques harmonisées et comparables. Donc, des cadres de qualité permettant de faire face à toutes ces demandes étaient nécessaires. Plusieurs cadres de qualité ont été élaborés et chacun comprend un certain nombre de dimensions de la qualité. L'exactitude et la pertinence ne sont que deux de ces dimensions.

Par exemple, le cadre élaboré par l'OCDE (2011) comprend huit dimensions, à savoir la pertinence, l'exactitude, l'actualité, la crédibilité, l'accessibilité, l'intelligibilité, la cohérence et la rentabilité (tableau 1). Des cadres similaires ont été établis par Statistique Canada (Statistique Canada 2002; Brackstone 1999), et par Statistics Sweden (Felme, Lyberg et Olsson 1976; Rosén et Elvers 1999). Le système statistique fédéral des États-Unis met depuis longtemps l'accent sur l'élément d'exactitude (U.S. Federal Committee on Statistical Methodology 2001), mais il apprécie certainement d'autres dimensions. Peut-être voit-il celles-ci comme étant de nature moins statistique, mais nécessitant néanmoins une part du budget total d'enquête. Le Fonds monétaire international (FMI) a élaboré un cadre qui diffère de

d'aider l'organisme statistique à s'améliorer. Les travaux de Deming (1986) ont influencé particulièrement les pratiques des enquêtes, car il insistait sur le rôle des statistiques dans l'amélioration de la qualité. Il faisait valoir vigoureusement l'idée que les statisticiens doivent diriger les travaux d'amélioration, puisqu'ils ont reçu une formation leur permettant de faire la distinction entre diverses formes de variation des processus. Selon lui, trop peu de chefs de file de la statistique conseillaient la haute direction des entreprises et il voulait que des statisticiens plus proactifs deviennent ce genre de chef de file. Il avait particulièrement à cœur de développer les idées de Shewhart au sujet des cartes de contrôle comme moyen de distinguer les divers types de variation, à savoir les variations ordinaires et les variations ayant une cause spéciale. Le cycle d'amélioration de Shewhart consistait à planifier-faire-contrôler-agir (*Plan-Do-Check-Act*) faisait également partie des réflexions de Deming sur la qualité (Shewhart 1939).

Naturellement, l'existence des principes de gestion remonte à des temps anciens. Juran (1995) donne une foule d'exemples de ceux qui étaient en place, par exemple, dans l'empire romain. Le savoir-faire des artisans et un système de guide en étaient les éléments fondamentaux. Des méthodes existaient pour choisir les matières premières et les fournisseurs. Les procédés étaient inspectés et améliorés. Les travailleurs étaient formés et motivés, et les clients obtenaient des garanties. Toutes ces caractéristiques se retrouvent encore dans les systèmes de gestion d'aujourd'hui. Les développeurs plus contemporains comprennent les cadres de la qualité ou les modèles d'excellence opérationnelle, tels que la gestion de la qualité totale (GTQ), les normes de l'Organisation internationale de normalisation (ISO), les critères du prix de qualité Malcolm Baldrige, le modèle d'excellence de la European Foundation for Quality Management (EFQM), les Six Sigma, les Lean Six Sigma et le tableau de bord prospectif (*Balanced Scorecard*). Ces modèles ne sont pas entièrement différents les uns des autres. Ils ont souvent en commun un ensemble de valeurs et de critères d'excellence. Ils représentaient plutôt une évolution naturelle que l'on peut constater dans toutes sortes de travaux.

Donc, on a assisté à l'adoption progressive des modèles de gestion de la qualité et des stratégies de qualité dans les organismes statistiques et à une fusion avec les concepts et les idées déjà appliqués par ces organismes. Mon calendrier personnel de cette évolution est le suivant (les lecteurs sont invités à produire des ensembles différents d'événements et de dates) :

1875 Taylor introduit ce qu'il appelle la gestion scientifique.

3. Principes de gestion de la qualité dans les organismes d'enquête

Malgré toutes ses limites, le cadre de l'erreur d'enquête totale présente des points forts assez convaincants. Il fournit une décomposition taxonomique des erreurs, sépare la variance du biais et l'observation de la non-observation, et définit les diverses étapes du processus d'enquête. Il sert de fondement conceptuel au domaine de la méthodologie d'enquête, les sous-domaines étant définis par leur structure d'erreur commeex. Enfin, il permet de cerner les lacunes dans la littérature sur la recherche, puisque toute typologie montrera que certaines étapes et processus sont plus « populaires » que d'autres. Il suffit pour s'en convaincre de comparer les portées respectives de la littérature sur la collecte des données et de celle sur le traitement des données. Il semble toutefois que le cadre de l'erreur d'enquête totale nécessite une extension dans des directions dont certaines avaient déjà été signalées il y a un demi-siècle. Nous avons besoin de directives afin de trouver un compromis entre la mesure de la taille des erreurs et l'obtention de processus davantage exempts d'erreur. La question que se pose Spencer (1985) est celle de savoir combien de ressources nous devons consacrer à la mesure par opposition à l'amélioration de la qualité. Nous avons également besoin de certaines directives quant à la façon d'intégrer des notions supplémentaires dans le cadre, afin qu'il devienne un cadre de la qualité d'enquête totale plutôt qu'un cadre de l'erreur d'enquête totale (Biemer 2010). Par exemple, si l'« adaptation à l'usage prévu » est le fondement conceptuel dominant, comment pouvons-nous lancer des travaux de recherche englobant la variation de l'erreur associée à différents usages ? Cet aspect est discuté à la section suivante.

Au cours des années 1980 et au début des années 1990, certains organismes statistiques faisaient face à de fortes pressions financières et, dans certains cas, étaient simultanément critiqués s'ils n'accordaient pas suffisamment d'attention aux besoins des utilisateurs. Les gouvernements de la Suède, de l'Australie, de la Nouvelle-Zélande et du Canada, de même que l'administration Clinton aux États-Unis souhaitaient vivement accroître l'efficacité de leurs systèmes statistiques respectifs, ainsi que l'influence exercée par les utilisateurs sur ces systèmes. Il était naturel pour ces organismes de s'inspirer des théories et méthodes de la gestion (Drucker 1985), tout spécialement ce que l'on appelle la gestion de la qualité (Juran et Gryna 1988). Grâce à cette nouvelle littérature, il était possible d'étudier le rôle du client, les problèmes de leadership, la notion d'amélioration continue de la qualité et les divers outils susceptibles

concernant les processus (appelé aujourd'hui paramètres), un document de conception et un plan B.

Les vues de Kish (1965) sur la conception des enquêtes différaient légèrement. Il favorisait les applications néobayésiennes en échantillonnage et la psychométrie prônée par certains collègues à l'Université du Michigan (Ericson 1969 ; Edwards, Lindman et Savage 1963). Par exemple, Kish aimait l'idée que des estimations au jugé des biais de mesure pourraient être combinées aux variances d'échantillonnage pour construire des estimations plus réalistes de l'erreur d'enquête totale. Quant au problème d'optimisation, il pensait qu'une approche polyvalente était économique : ment favorable pour les enquêtes, mais qu'il pourrait être difficile de décider sur quoi fonder le plan de sondage. Si l'on arrive à désigner une statistique principale, celle-ci peut à elle seule déterminer le plan de sondage et s'il existe un petit nombre de ces statistiques, il est possible d'opter pour un plan de sondage de compromis ; par contre, si les statistiques sont trop disparates, il pourrait n'exister aucun plan de sondage raisonnable. Kish insiste aussi sur la nécessité d'obtenir des renseignements sur le plan de sondage au moyen d'enquêtes pilotes et de prétests afin de prendre plus facilement les décisions concernant ce plan. Il a constaté que le plan de sondage et les mesures pouvaient varier considérablement selon l'environnement, tandis que l'échantillonnage changeait moins. Il pourrait s'agir de l'une des raisons pour lesquelles l'échantillonnage peut être classé facilement parmi les théories et méthodes statistiques classiques, alors qu'il est plus difficile d'insérer le processus d'enquête dans une discipline particulière (Frankel et King 1996 dans leur entrevue avec Kish).

Comme les autres témoins de la conception des enquêtes, Kish insistait sur l'importance d'un faible biais, mais appréciait le fait que la réduction du terme de biais pourrait accroître l'erreur totale. Il avait à cœur d'arriver à un équilibre raisonnable entre les diverses sources d'erreur et la façon dont les structures d'erreur variaient sous diverses options du plan de sondage. Comme Hansen et ses collègues, Kish pensait que les renseignements pertinents devaient être enregistrés simultanément durant la mise en œuvre (de nouveau nous voyons le parallèle avec les paramètres), Hansen et ses collègues s'inquiétaient vraiment de l'application de contrôles excessifs, mais inadéquats. Ils se sont rendu compte que certains contrôles devaient peut-être être relâchés en raison des améliorations limitées qui en découlait et que le degré d'amélioration des estimations devait être vérifié avant de procéder à tout relâchement des contrôles. Ils ont également suggéré que l'on devrait peut-être compromettre la pertinence pour obtenir des mesures contrôlables ou s'abstenir de procéder à l'enquête. Tant Hansen et ses collègues que Kish défendaient vivement l'idée de mettre fin à la pratique voulant que

l'erreur d'échantillonnage soit la seule erreur d'enquête mesurée.

Un examen de la situation actuelle porte à conclure que l'on ne dispose toujours pas d'une formule de conception des enquêtes. Il n'existe pour ainsi dire aucun manuel de planification, et la littérature sur la conception des enquêtes est par conséquent peu abondante, de même que celle sur les coûts (Groves 1989 est une exception). Aucune formule de conception n'est en vue. Depuis l'élaboration du modèle d'enquête du U.S. Census Bureau, plusieurs variantes ont fait leur apparition, certaines d'entre elles assez compliquées (Groves et Lyberg 2010). Une caractéristique commune est le fait qu'elles ont tendance à être incomplètes, c'est-à-dire qu'elles ne tiennent pas compte de toutes les sources d'erreur. Sur le plan statistique, l'attention se concentre surtout sur les composantes de la variance, en particulier la variance de l'erreur de mesure. Un certain nombre d'autres faiblesses sont associées au concept de l'erreur d'enquête totale. En premier lieu, la perspective des utilisateurs fait défaut et une vaste majorité d'entre eux ne sont pas à même de mettre en doute l'exactitude des données ni d'en discuter. Les structures et les interactions complexes des erreurs n'incitent pas les contrôles extérieurs et les contacts avec les utilisateurs ont souvent tendance à porter sur des questions moins techniques, telles que l'actualité, la comparabilité et le coût des données. Les utilisateurs ne sont pas vraiment au courant des niveaux réels d'exactitude et nous en savons fort peu quant à la façon dont ils perçoivent l'information sur les erreurs et comment y donner suite.

Comme l'a fait remarquer Biemer (2001), il existe un manque de mesures systématiques des composantes de l'EQM dans les organismes statistiques. Plusieurs bonnes raisons sont à l'origine de cette situation. À la complexité, qui a déjà été mentionnée, nous pouvons ajouter des facteurs tels que les coûts, le fait qu'il est presque impossible de publier ce genre d'information au moment où les données sont diffusées et le fait qu'il n'existe aucune mesure de l'erreur totale qui tient compte de toutes les sources d'erreur, faute d'une méthodologie appropriée ou parce qu'il est impossible d'exprimer certaines erreurs. Groves et Lyberg (2010) énumèrent certaines autres faiblesses du paradigme de l'erreur d'enquête totale. Par exemple, nous devons en savoir davantage sur l'interaction entre les variances et les biais. Il se peut qu'un accroissement de la simple variance de réponse aille de pair avec une réduction du biais de réponse, disons, quand nous comparons le mode d'interview à des options d'autoadministration du questionnaire. Récemment, West et Olson (2010) ont montré que la variance due à l'intervieweur peut résulter non seulement de l'effet individuel des intervieweurs sur les réponses recueillies dans le cadre de leurs tâches, mais aussi du fait que les intervieweurs réussissent individuellement à obtenir

provenant de chaque source connue d'erreur. Groves et Lyberg (2010) résument la situation du paradigme dans la pratique passée et contemporaine des enquêtes.

L'idée selon laquelle les enquêtes doivent être conçues en tenant compte de toutes les sources d'erreur émane des premiers témoins du domaine. Morris Hansen, Bill Hurwitz, Joe Waksberg, Leon Pritzker, Ed Deming et d'autres au U.S. Census Bureau, Leslie Kish à l'Université du Michigan, P.C. Mahalanobis à l'Institut statistique de l'Inde, et Tore Dalenius, à l'Université de Stockholm, étaient parmi les chefs de file de la recherche sur les enquêtes, mettant l'accent sur les erreurs et l'optimisation du plan de sondage. Ils étaient préoccupés par les limites inhérentes à la théorie de l'échantillonnage, car les erreurs non dues à l'échantillonnage risquaient de faire s'effondrer la théorie. Très pragmatiques, ils réfléchissaient beaucoup à la façon d'équilibrer les erreurs et aux coûts qu'entraîne leur traitement. Voyant des similitudes entre une chaîne de montage d'usine (Deming et Geoffrey 1941) et la mise en œuvre de certains processus d'enquête, certains d'entre eux ont introduit des méthodes de contrôle tirées d'applications industrielles.

Dalenius (1967) s'est rendu compte qu'il n'existait pas encore de « formule de conception » pouvant fournir une solution optimale au problème. L'approche adoptée par Dalenius ainsi que par Hansen, Hurwitz et Pritzker (1967) consistait à minimiser tous les biais et à opter pour un scénario de variance minimale, pour que la variance devienne une approximation de l'EQM. Cela était censé se faire au moyen de schémas de vérification intense pour les productions en cours et d'études d'évaluation d'assez grande portée pour les futures productions. En 1969, inspiré par Hansen, Dalenius a présenté une communication portant sur la conception globale des enquêtes (*total survey design*), où le terme « totale » traduisait l'idée de prendre en compte toutes les sources d'erreur. Hansen, Hurwitz, Marks et Mauldin (1951), Hansen, Hurwitz et Bershad (1961), et Hansen, Hurwitz et Pritzker (1964) ont élaboré le modèle d'enquête du U.S. Census Bureau qui tenait compte de l'effet des intervieweurs, des codeurs, des vérificateurs et des chefs d'équipe, et permettait d'estimer leur contribution à l'erreur d'enquête totale. Ces schémas d'estimation, variations de la répétition et de l'interpénétration. L'hypothèse était que l'estimation du biais était traitée par comparaison des estimations obtenues d'après les opérations ordinaires à celles obtenues au moyen des procédures privilégiées (qui ne pouvaient pas être utilisées à grande échelle pour des raisons financières, administratives ou pratiques). Aujourd'hui, ce genre d'approche est considéré comme étant la « norme de référence » (*gold standard*).

Bien que Hansen, Dalenius et d'autres aient été les premiers à préconiser la conception globale des enquêtes, il était rare que les utilisateurs participent directement à la détermination finale des exigences concernant l'enquête. Assez souvent, un agent, un administrateur ou un statisticien jouait le rôle de spécialiste du domaine. Il y a plusieurs décennies, c'est comme cela que nous pensions aux utilisateurs. Leurs opinions comptaient, mais ils ne participaient pas vraiment aux prises de décisions. Cependant, au fond de nous-mêmes, nous savions qu'il ne s'agissait peut-être pas d'un modèle parfait et, à la fin des années 1970, Statistics Sweden a publié une brochure interne intitulée « Que faire si un client se présente à notre porte ».

L'approche fondamentalement de conception proposée par Hansen, Dalenius et d'autres comprenait un certain nombre d'étapes, dont :

- la spécification d'un objectif idéal d'enquête ;
- l'analyse de la situation de l'enquête quant aux ressources en matière de budget, de méthodologie et d'information ;
- l'élaboration d'un petit nombre d'options de plan d'enquête ;
- l'évaluation des diverses options en se basant sur les déterminations préliminaires connexes de l'EQM et des coûts ;
- le choix de l'une des options ou d'une modification de l'une d'elles, ou la décision de ne pas procéder du tout à l'essai de faisabilité, un système de signalisation

soutenait que, jusqu'à ce que le but soit énoncé, il n'existait aucune bonne ou mauvaise façon d'entreprendre une enquête. Certains autres statisticiens ont fait des déclarations comparables. En fait, c'est vraiment l'utilisateur qui était caché derrière des termes tels que « problème lié au domaine spécialisé », « but de l'étude » ou « fonctions clés d'un système statistique ».

Même aujourd'hui, les concepts d'enquête et de qualité sont vagues. Comme l'on souligne Morgansstein et Marker (1997), les définitions variées de la qualité nuisent aux travaux d'amélioration, de sorte que nous devons au moins essayer de faire la distinction entre les différentes définitions afin de déterminer à quoi elles servent. L'un des défis importants que la qualité est une fonction directe de l'« adaptation à l'usage prévu ». En fait, déjà en 1944, Deming avait utilisé la phrase « fitness for purpose » (adaptation au but pour suivre), non pas pour définir la qualité, mais plutôt pour expliquer ce qui faisait la réussite d'un produit d'enquête.

Longtemps, la notion de « bonne » qualité était implicitement équivalente à une faible erreur quadratique

moyenne (EQM), ce qui signifie que les données doivent être exactes et que l'exactitude d'une estimation peut être mesurée par l'EQM, qui est la somme de la variance et du carré du biais. Nous avons constaté que les statistiques fondées sur des sondages doivent aussi être utiles, ce qui a été désigné plus tard par le terme « pertinentes ». Nombre des dimensions actuelles de la qualité ne représentaient pas vraiment un sujet de préoccupation à l'époque. En outre, les utilisateurs étaient habitués à ce que la réalisation des enquêtes prenne du temps ; l'actualité des données était certes à l'ordre du jour, mais pas aussi explicitement qu'aujourd'hui. Le traitement des données d'un recensement prenait des années. Les utilisateurs étaient accoutumés à une technologie qui ne permettait d'offrir que des formes assez simples d'accèsibilité. Donc, il était naturel pour les utilisateurs et les producteurs de faire en sorte avant tout que le problème statistique concorde raisonnablement avec le problème du secteur spécialisé et que l'EQM soit maintenue à un niveau acceptable. Cette EQM était, et est encore, équivalente dans de nombreux cas à la variance seulement, sans ajout d'un terme de carré du biais.

Avant de poursuivre, définissons ce qu'est une « enquête ». Une enquête est une étude statistique conçue pour mesurer les caractéristiques de la population afin de pouvoir estimer les paramètres de cette dernière. La proportion de chômeurs à un moment donné dans une population de personnes, ou le revenu total d'une entreprise ou d'un secteur d'activité durant une période donnée sont deux exemples de paramètres. Une enquête peut être définie comme une liste de conditions préalables (Dalenius 1985a). Selon Dalenius, une étude peut être catégorisée comme une

enquête si les conditions préalables qui suivent sont satisfaites :

1. l'étude concerne un ensemble d'objets constituant une population ;
2. la population étudiée possède une ou plusieurs propriétés mesurables ;
3. le but de l'étude est de décrire la population au moyen d'un ou de plusieurs paramètres définis en fonction des propriétés mesurables, ce qui nécessite l'observation (d'un échantillon) de la population ;
4. pour arriver à observer la population, une base de sondage est nécessaire ;
5. un échantillon d'objets est sélectionné à partir de la base de sondage conformément à un plan d'échantillonnage qui spécifie un mécanisme probabiliste et une taille d'échantillon n (où n pourrait être égal à N , la taille de la population) ;
6. des observations sont faites sur l'échantillon conformément à un procédé de mesure (c'est-à-dire une méthode de mesure et une prescription concernant son utilisation) ;
7. un processus d'estimation fondé sur les mesures est appliqué pour calculer des estimations des paramètres lorsque l'on fait une inférence au sujet de la population étudiée d'après l'échantillon.

Cette définition énumère implicitement les sources particulières d'erreur présentes dans les travaux d'enquête. Pour chaque source, il existe un certain nombre de méthodes qui en minimisent les effets, mais mesurent également leur grandeur (Biemer et Lyberg 2003 ; Groves, Fowler, Couper, Lepkowski, Singer et Tourangeau 2009).

Les écarts par rapport à la définition reflètent des défauts de qualité. En outre, les écarts de ce genre sont fréquents. Dans certains plans de sondage, les probabilités de sélection sont inconnues ou l'estimateur de la variance choisi n'est pas nécessairement celui qui convient le mieux, étant donné le plan utilisé. Le fait que ces défauts posent problème ou non dépend du but de l'enquête.

2.2 Les composantes du paradigme de l'erreur

d'enquête totale

Le paradigme de l'erreur d'enquête totale est un cadre théorique utilisé pour optimiser les enquêtes en minimisant la grandeur cumulée des erreurs provenant de toutes les sources, étant donné des contraintes budgétaires. En pratique, cela signifie que nous voulons minimiser l'erreur quadratique moyenne de certaines estimations fondées sur les données d'enquête, à savoir celles que les principales parties prenantes jugent les plus importantes. L'erreur quadratique moyenne, qui est la mesure utilisée le plus fréquemment pour évaluer le travail d'enquête, est égale à la

Bureau, où avait lieu à l'époque une grande partie des activités de recherche appliquée et d'élaboration de nouvelles théories. L'un des résultats remarquables des travaux du Census Bureau a été la production d'un manuel en deux volumes, sur la théorie et les méthodes d'échantillonnage (Hansen, Hurwitz et Madow 1953). En fait, les progrès en théorie de l'échantillonnage étaient si importants à ce moment-là que Stephan (1948) a jugé bon de rédiger un article sur l'histoire des méthodes modernes d'échan-

tillonnage.

Très tôt, on a reconnu qu'il pouvait exister d'autres erreurs d'enquête que celles attribuées à l'échantillonnage. Il existait des écrits sur les effets du libellé des questions, dont celui de Muscio (1917). La recherche sur la conception des questionnaires était assez intensive durant les années 1940. Mahalanobis (1946) s'est attaqué aux problèmes résultants des erreurs introduites par les enquêteurs sur le terrain chargés de recueillir les données agricoles en Inde, ce qui a donné une méthode d'estimation de ces erreurs. Cette méthode, appelée « interpendentation », peut être utilisée pour estimer ce que l'on appelle les variances corrélées introduites par les intervieweurs, les vérificateurs, les codeurs et les personnes qui supervisent ces groupes. Les sources d'erreur les plus importantes étaient certainement déjà connues autour de 1950. Deming a dressé une liste des sources d'erreur (1944) qui constitue la première typologie publiée des erreurs d'enquête, et Hansen et Hurwitz (1946) ont discuté du sous-échantillonnage des non-répondants pour essayer de fournir des estimations sans biais dans une situation présentant une non-réponse initiale. Cependant, sur le plan de la méthodologie, l'accent avait été mis jusque-là sur l'élaboration de la théorie de l'échantillonnage, ce qui est assez compréhensible. Il était en effet très important de pouvoir montrer qu'il était possible de réaliser des enquêtes en s'appuyant sur l'échantillonnage, et ce, dans diverses conditions. En 1950, il avait été démonté de manière assez satisfaisante que cela était effectivement faisable. Donc, il était temps de passer à d'autres questions et aux peau-

finements.

Au début, l'emploi du terme qualité était limité avant tout au contrôle de la qualité, parfois au contrôle de la qualité des opérations d'enquête. Souvent, le contrôle de la qualité se résumait à la vérification et (ou) à l'estimation de la grandeur de l'erreur pour diverses opérations. On savait que les statistiques étaient affectées par d'autres erreurs que celles émanant de l'échantillonnage, mais la façon, liée à la qualité des processus, de réduire systématiquement ces erreurs et biais restait encore à établir (Deming 1944 ; Hansen et Steinberg 1956).

Il y a 60 ans d'ici, l'utilisateur était un joueur plutôt obscur, même si les éminents concepteurs des techniques d'enquête ne l'ignoraient pas du tout. Ainsi, Deming (1950)

qu'appartiennent la plupart des exemples de qualité des enquêtes que j'expose.

La présentation de l'article est la suivante. La section 2 traite du paradigme de l'erreur d'enquête totale, y compris les typologies de l'erreur, le traitement des erreurs et la conception des enquêtes en tenant compte de toutes les sources d'erreur. La section 3 porte sur les approches de gestion de la qualité qui ont eu un effet important sur les organismes d'enquête depuis le début des années 1990. Cet effet s'est manifesté par des méthodes et des approches telles que la prise en compte de l'utilisateur ou du client, la discussion des coûts et des risques dans le cadre de la recherche sur les enquêtes, et la nécessité pour les organismes de continuer à s'améliorer. La section 4 fournit des exemples de projets d'amélioration de la qualité entrepris par les organismes d'enquête. La section 5 traite des difficultés que pose la mesure, directe ou indirecte, de la qualité au moyen d'indicateurs. Est également abordée la façon dont ces mesures doivent être communiquées aux utilisateurs ou aux clients. Enfin, la section 6 offre certaines réflexions quant à la manière dont les pratiques d'enquête *doivent* évoluer afin de mieux répondre aux besoins des utilisateurs. La dernière section est réservée à la bibliographie.

2. Le paradigme de l'erreur d'enquête totale

2.1 Bref historique de l'échantillonnage

Un certain nombre d'articles décrivent l'élaboration des premières méthodes d'échantillonnage. On constate dans ces premiers travaux une reconnaissance implicite ou explicite des problèmes de qualité, même s'ils sont masqués sous des termes tels que « erreurs » et « utilité de l'enquête » (Deming 1944). Les aperçus historiques que l'on trouve, par exemple, dans Kish (1995), Fienberg et Tanur (1996), et O'Muircheartaigh (1997) insistent tous sur le fait que, jusqu'à 1950, on a assisté au plein essor de la théorie de l'échantillonnage. Dans les années 1920, l'Institut international de statistique a accepté de promouvoir les idées sur l'échantillonnage pour être mesurée en calculant la variance de l'estimateur. Bill Cochran, Frank Yates, Ed Deming, Morris Hansen et bien d'autres ont perfectionné les concepts de la théorie de l'échantillonnage. Hansen a dirigé un groupe de recherche au U.S. Census

La qualité des enquêtes

Lars Lyberg¹

Résumé

La qualité des enquêtes est un concept multidimensionnel issu de deux démarches de développement distinctes. La première démarche suit le paradigme de l'erreur d'enquête totale, qui repose sur quatre piliers dont émanent les principes qui guident la conception de l'enquête, sa mise en œuvre, son évaluation et l'analyse des données. Nous devons concevoir les enquêtes de façon que l'erreur quadratique moyenne d'une estimation soit minimisée compte tenu du budget et d'autres contraintes. Il est important de tenir compte de toutes les sources communes d'erreur, de surveiller les principales d'entre elles durant la mise en œuvre, d'évaluer périodiquement les principales sources d'erreur et les combinaisons de ces sources après l'achèvement de l'enquête, et d'étudier les effets des erreurs sur l'analyse des données. Dans ce contexte, on peut mesurer la qualité d'une enquête par l'erreur quadratique moyenne, la contrôler par des observations faites durant la mise en œuvre et l'améliorer par des études d'évaluation. Le paradigme possède des points forts et des points faibles. L'un des points forts tient au fait que la recherche peut être définie en fonction des sources d'erreur et l'un des points faibles, au fait que la plupart des évaluations de l'erreur d'enquête totale sont incomplètes, en ce sens qu'il est impossible d'inclure les effets de toutes les sources. La deuxième démarche est influencée par des idées empruntées aux sciences de la gestion de la qualité. Ces sciences ont pour objet de permettre aux entreprises d'exceller dans la fourniture de produits et de services en se concentrant sur leurs clients et sur la concurrence. Ces idées ont eu une très grande influence sur de nombreux organismes statistiques. Elles ont notamment amené les fournisseurs de données à recommander qu'un produit de qualité ne peut pas être obtenu si la qualité des processus sous-jacents n'est pas suffisante et que des processus de qualité suffisante ne peuvent pas être obtenus sans une bonne qualité organisationnelle. Ces divers niveaux peuvent être contrôlés et évalués au moyen d'ententes sur le niveau de services, de sondages auprès des clients, d'analyses des paradoxes en recourant au contrôle statistique des processus et d'évaluations organisationnelles en se servant de modèles d'excellence opérationnelle ou d'autres ensembles de critères. À tous les niveaux, on peut rehausser la qualité en lançant des projets d'amélioration choisis selon des fonctions de priorité. L'objectif ultime de ces projets d'amélioration est que les processus concernés s'approchent progressivement d'un état où ils sont exempts d'erreur. Naturellement, il pourrait s'agir d'un objectif impossible à atteindre, mais auquel il faut tenir de parvenir. Il n'est pas raisonnable d'espérer obtenir des mesures continues de l'erreur d'enquête totale en se servant de l'erreur quadratique moyenne. Au lieu de cela, on peut espérer qu'une amélioration continue de la qualité par l'application des idées des sciences de la gestion ainsi que des méthodes statistiques permettra de minimiser les biais et d'autres problèmes que posent les processus d'enquête, afin que la variance devienne une approximation de l'erreur quadratique moyenne. Si nous y arrivons, nous aurons fait coïncider approximativement les deux démarches de développement.

Mots clés : Gestion de la qualité ; erreur d'enquête totale ; cadre de la qualité ; erreur quadratique moyenne ; variabilité des processus ; contrôle statistique des processus ; utilisateurs des données d'enquête.

1. Introduction

Le présent article a été rédigé en reconnaissance des apports uniques et du leadership de Joe Waksberg dans le domaine des techniques d'enquête. J'ai pris connaissance des travaux de Joe pour la première fois en lisant son article sur les erreurs de réponse dans les enquêtes sur les dépenses, rédigé en collaboration avec John Neter (Neter et Waksberg 1964). Entre autres, cet article m'a fait découvrir le phénomène cognitif appelé télescopage. Plus tard, j'ai eu l'occasion de travailler avec Joe à la préparation de la première conférence et monographie sur les méthodes d'enquête téléphonique en tant que membre du comité de rédaction (Groves, Biemer, Lyberg, Massey, Nicholls et Waksberg 1988). Nous avons également collaboré à la préparation de nombreuses conférences Morris Hansen dont les exposés ont été publiés dans le *Journal of Official Statistics* (JOS) durant mon mandat de rédacteur en chef. Joe lui-même a donné la sixième conférence, qui a été publiée dans le JOS

(Waksberg 1998). Joe était un fantastique chef de file et c'est pour moi un grand honneur d'avoir été invité à rédiger cet article sur la qualité des enquêtes, sujet qui le préoccupait beaucoup. Bon nombre de mes amis m'ont fait part de leurs opinions ou m'ont envoyé de la documentation en prévision du présent article. Je remercie tout spécialement Paul Biemer, Dan Kasprzyk, Fritz Scheuren, Dennis Trewin et Maria Bohata de leur aide. La qualité des enquêtes est un concept vague, quoiqu'intuitif, ayant de nombreuses significations. Dans le présent article, je discute de certaines observations qui ont trait à l'élaboration et au traitement du concept au cours des soixante-dix dernières années et, dans le cas de certains développements, il m'est même possible de remonter à des origines encore plus lointaines. Toutefois, ma discussion porte en majeure partie sur les questions qui se posent aujourd'hui dans les organismes statistiques gouvernementaux. C'est au domaine de la statistique officielle

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg.

Veuillez consulter la section avis à la fin de la revue pour des informations sur le processus de nomination et de sélection du prix Waksberg 2014.

Ce numéro de *Techniques d'enquête* commence par le douzième article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé d'Elizabeth A. Martin (présidente), Mary E. Thompson, Steve Heeringa et J.N.K. Rao, d'avoir choisi Lars Lyberg comme auteur de l'article du prix Waksberg de cette année.

Communication sollicitée pour le prix Waksberg 2012

Auteur : Lars Lyberg

Lars Lyberg, Ph. D., est l'ancien chef du Département de la recherche-développement de Statistique Suède et est actuellement professeur émérite au Département de statistique de l'Université de Stockholm. Il a fondé le *Journal of Official Statistics (JOS)* et y a été rédacteur en chef pendant 25 ans. Il est rédacteur en chef de *Survey Measurement and Process Quality* (Wiley 1997) et corédacteur de *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (Wiley 2010), *Telephone Survey Methodology* (Wiley 1988) et *Measurement Errors in Surveys* (Wiley 1991). Il est coauteur de *Introduction to Survey Quality* (Wiley 2003). Il a présidé le groupe de travail sur la qualité du Système statistique européen ainsi que le comité organisateur de la première Conférence européenne sur la qualité des statistiques officielles, Q2001. Il a déjà présidé l'AISE ainsi que la Section des méthodes d'enquête de l'American Statistical Association. Il est membre de cette dernière association et de la Royal Statistical Society.



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'«American National Standard for Information Sciences» — «Permanence of Paper for Printed Library Materials», ANSI Z39.48 - 1984.

Techniques d'enquête

Une revue éditée par Statistique Canada
Volume 38, numéro 2, décembre 2012

Table des matières

Article sollicité Waksberg

Lars Lyberg	La qualité des enquêtes	115
-------------	-------------------------------	-----

Articles réguliers

Jaqueline Garcia-Vi et Ulrike Grote	Collecte de données : expérience et leçons apprises au chapitre des questions de nature délicate dans une région éloignée de culture de la coca au Pérou	143
-------------------------------------	--	-----

Jun Shao, Martin Klein et Jing Xu	Imputation pour la non-réponse non monotone dans le <i>Survey of Industrial Research and Development</i>	157
Jae Kwang Kim et Minsun Kim Riddles	Théorie concernant les estimateurs ajustés sur le score de propension dans les sondages	171

Ian Plewis, Sosihenes Ketende et Lisa Calderwood	Évaluation de l'exactitude des modèles de propension à répondre dans les études longitudinales	181
--	--	-----

Sarat C. Dass, Tapabratra Maity, Hao Ren et Samiran Sinha	Estimation des intervalles de confiance des paramètres de petit domaine avec rétrécissement des moyennes et des variances	187
---	---	-----

Dan Liao et Richard Valliant	Indices de conditionnement et décompositions des variances pour le diagnostic de la collinéarité dans l'analyse de données d'enquête au moyen de modèles linéaires	205
Qixuan Chen, Michael R. Elliott et Roderick J.A. Little	Inférence bayésienne pour les quantiles de population finie sous échantillonnage avec probabilités inégales	221

Communications brèves

Satkarar K. Kinney	Imputation multiple dans le cas de données de recensement	235
--------------------	---	-----

Avertissement	239
Corrigendum	240
Remerciements	241
Annonces	243
Autres revues	245

Techniques d'enquête est répertoriée dans *The ISI Web of knowledge (Web of science)*, *The Survey Statistician*, *Statistical Theory and Methods Abstracts* et *SRM Database of Social Research Methodology*, *Erasmus University*. On peut en trouver les références dans *Current Index to Statistics*, et *Journal Contents in Qualitative Methods*. La revue est également citée par *SCOPUS* sur les bases de données *Elsevier Bibliographic Databases*.

COMITÉ DE DIRECTION

Président

J. Kovar

Anciens présidents

D. Royce (2006-2009)

G.J. Brackstone (1986-2005)

R. Plalek (1975-1986)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.A. Hidiroglou, *Statistique Canada*

chef délégué

H. Mantel, *Statistique Canada*

Rédacteurs associés

J.-F. Beaumont, *Statistique Canada*

J. van den Brakel, *Statistics Netherlands*

J.M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

R. Chambers, *Centre for Statistical and Survey Methodology*

J.T. Elinga, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistique Canada*

D. Haziza, *Université de Montréal*

B. Hühlig, *University of Applied Sciences Northwestern Switzerland*

D. Judkins, *Westat Inc.*

J.K. Kim, *Iowa State University*

P.S. Kott, *RTI International*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallée, *Statistique Canada*

P. Lynn, *University of Essex*

D.J. Males, *National Center for Health Statistics*

G. Nathan, *Hebrew University*

Rédacteurs adjoints

Y. You, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyses de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou d'appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféablement en Word au rédacteur en chef, (tre@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Turney, Ottawa, Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentes dans la revue ou sur le site web (www.statcan.gc.ca).

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 \$ CA (6 \$ x 2 exemplaires); autres pays, 20 \$ CA (10 \$ x 2 exemplaires). L'Association des statisticiens de l'Amérique Statistician Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statisticienne du Canada et l'Association des statisticiens et statisticiennes du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.

Techniques d'enquête

Une revue éditée par Statistique Canada

Décembre 2012 • Volume 38 • Numéro 2

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2012

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division de la gestion de l'information, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Décembre 2012

N° 12-001-XPB au catalogue

Périodicité : semestrielle

ISSN 0714-0045

Ottawa

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements
Service national d'appareils de télécommunications pour les malentendants
Télécopieur
1-800-263-1136
1-800-363-7629
1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements
Télécopieur
1-613-951-8116
1-613-951-0581

Programme des services de dépôt

Service de renseignements
Télécopieur
1-800-635-7943
1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit n° 12-000-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Ce produit est aussi disponible en version imprimée standard au prix de 30 \$CAN l'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste
- Finances
Statistique Canada
Immeuble R.-H.-Coats, 6^e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Techniques d'enquête

N° 12-001-XPB au catalogue

Une revue

éditée

par Statistique Canada

Décembre 2012

•

Volume 38

•

Numéro 2



Statistique
Canada

Statistics
Canada

Canada

